UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE CIENCIAS MATEMÁTICAS

E.A.P. DE ESTADÍSTICA

Estimación de observaciones faltantes mediante la media aritmética y su efecto en el análisis de componentes principales

TESIS

Para optar el Título Profesional de Licenciada en Estadística

AUTOR

Eduviges Cleofé de los Ríos Tello

ASESOR

Doris Gómez Ticerán

Lima - Perú

2011

A Dios por la fortaleza

Que me da y al amor

Incondicional de mis padres:

Víctor Alejandro De los Ríos Cárdenas

Elena Tello Adrían; hnos. Víctor y Ana;

Tíos: Grocio y Segundino.

AGRADECIMIENTO

El presente trabajo de tesis tiene por objetivo presentar las demostraciones del uso de la media aritmética, de los datos observados, para estimar observaciones faltantes de la muestra; luego evaluar el efecto de dicha manera de estimar las observaciones faltantes en el Análisis de Componentes Principales.

La elaboración de este trabajo necesitó de la cooperación de algunas personas, por lo que quiero expresarles mi agradecimiento, pues me ayudaron de manera desinteresada y muchas de ellas han sido un soporte contundente para mí hasta el día de hoy.

A mis padres y hermanos, porque gracias a sus esfuerzos he podido terminar mis estudios y consiguientemente esta tesis.

A mi asesora la Dra. Doris Gómez Ticerán, por motivarme y guiarme en este proyecto, por su paciencia, consejos y enseñanzas.

Al Magister Teodoro Sulca Paredes, por haberme brindado la orientación matemática que permitió culminar el presente trabajo.

A los profesores de la Escuela Académico Profesional de Estadística, por ser los artífices de mi formación académico-profesional.

A mis amigos, por compartir grandes momentos en la vida universitaria y a todas las personas que me apoyaron para llevar a cabo este proyecto.

ESTIMACIÓN DE OBSERVACIONES FALTANTES MEDIANTE LA MEDIA ARITMETICA Y SU EFECTO EN EL ANALISIS DE COMPONENTES PRINCIPALES

EDUVIGES CLEOFÉ DE LOS RIOS TELLO

Tesis presentada a consideración del Cuerpo Docente de la Facultad de Ciencias Matemáticas, de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para optar el Título Profesional de Licenciada en Estadística

Aprobada por:

Mg.Olga Lidia Solano Dávila _____
Presidenta

Mg.Ysabel Adriazola Cruz _____
Miembro

Dra. Doris Gómez Ticerán _____
Miembro - Asesor

Lima – Perú

2011

FICHA CATALOGRÁFICA

DE LOS RIOS TELLO EDUVIGES CLEOFÉ

Estimación De Observaciones Faltantes Mediante La Media Aritmética y su efecto en el Análisis de Componentes Principales

101p, 30 cm. (UNMSM, Licenciada en Estadística, 2011)

Tesis, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas. Estadística.

I. UNMSM / F. de C.M. II. Estimación de Observaciones Faltantes mediante la media y su efecto en el análisis de componentes principales

RESUMEN

Se demuestra que la estimación de los datos faltantes mediante la media aritmética de los datos conocidos, no tiene efecto y/o modificación en el valor de la media aritmética de todos los datos, incluidos los datos faltantes previamente estimados por la media aritmética de los datos conocidos y sus consecuencias en el análisis de componentes Principales.

Se demuestra que las componentes principales usando la matriz de covarianzas y correlaciones no han sido alterados respecto a su valor original con todos los datos conocidos y con datos faltantes estimados mediante las medias.

El software que se utilizó para el análisis de datos fue el R.

Palabras Claves:

Estimadores puntuales. Media. Varianza. Componentes Principales. Datos faltantes.

SUMMARY

We show that the estimation of missing data by the arithmetic average of the known data, has no effect and / or change in the value of the arithmetic mean of all data, including previously missing data estimated by the arithmetic mean of the data known and its impact on the Analysis component Principal .

We show that the major components using the covariance and correlation matrix are not altered compared to its original value with all the known facts and with missing data estimated by the averages.

The software used for data analysis was the R.

Key Words:

Point Estimators.Mean.Variance.Principal Component.Data missing

ÍNDICE

INTRODUCCIÓN

En las encuestas de hogares la falta de respuesta de las personas encuestadas se asocia a diversas causas: A la fatiga del informante, al desconocimiento de la información solicitada, al rechazo de las personas a informar acerca de temas sensibles, a la negativa de los hogares a participar en la investigación, así como a problemas asociados a la calidad del marco de muestreo.

A pesar de que un cuestionario sea considerado correcto, la realidad indica que frecuentemente los archivos contienen observaciones "aberrantes" o "poco probables", y existen situaciones en que debido a los objetivos de la investigación, deliberadamente se omite información de personas que no forman parte de la población de estudio.

Por ejemplo, en una encuesta de demografía y salud, la fecundidad tradicionalmente se estudia en las mujeres de 14 años y más, y por definición, en la base de datos existirán registros sin información para las personas que no forman parte de la población objetivo. La mayoría de los casos se le asigna un código "." Que comúnmente se asocia con una información faltante.

Los datos faltantes forman parte de un conjunto de observaciones con características especiales que incluyen a los datos agrupados, agregados, redondeados, censurados o truncados; es decir, a datos con información especial (Heitjan y Rubin, 1991)

En los últimos cien años ha sido enorme el desarrollo de métodos estadísticos para estimar datos incompletos. El algoritmo EM y sus extensiones, la imputación y el método de Monte Carlo proporcionan un conjunto de herramientas flexibles y confiables para hacer inferencia en una gran cantidad de problemas de datos faltantes. Sin embargo, en términos prácticos, estos desarrollos han tenido pequeño impacto en la forma de analizar datos, en presencia de valores perdidos u observaciones faltantes (Schafer, 1999).

Existen varios tipos de observaciones faltantes (Mandevile; 2010).

Faltantes completamente al azar (Missing Completely at Random MCAR), se refiere a los datos, cuando el proceso de pérdida no depende de las otras variables explicativas en el conjunto de datos. Con MCAR, cualquier observación tiene la misma probabilidad de perderse, esto significa que los datos se recolectan al azar, y no dependen de ninguna otra variable en el conjunto de datos. Ejemplo, las observaciones que pueden perderse porque el equipo no funcionaba correctamente, la gente se enfermó o los datos no se capturaron correctamente.

Faltantes al azar (Missing at Random, MAR). MAR significa que las observaciones faltantes están condicionadas por otras variables explicativas en el conjunto de datos, aunque no con la variable respuesta.(Schafer, 1999). Por ejemplo, las personas que sufren depresión podrían estar menos inclinadas a reportar sus ingresos económicos, por lo que reportarlos se relaciona con su depresión. Las personas deprimidas pueden tener ingresos más bajos y, por lo tanto, una alta tasa de observaciones perdidas, el ingreso medio calculado podría ser menor de lo que es en un grupo sin observaciones perdidas.

Faltantes no al azar (Missing not at random, MNAR).Si los datos no son MAR o MCAR entonces se denominan MNAR. Por ejemplo, si se está estudiando la salud mental de personas diagnosticadas como deprimidas y éstas son menos propensas que otras a informar sobre su estado mental, no hay una pérdida de observaciones al azar. Lo mismo sucede cuando las personas con bajos ingresos son menos propensas a reportar sus ingresos.

Con respecto a los métodos de estimación de observaciones faltantes, las formas mas generales de abordarlas son:

Supresión o análisis de los casos completos. Cuando el número de casos con observaciones perdidas es pequeño (menosdel 5% en muestras grandes) es común suprimir esos casos, y ejecutar el análisis con las observaciones que quedan. Aunque este método puede resultar en una disminución del tamaño de la muestra, es uno de lo más utilizados.

Sustitución por la media. Un antiguo procedimiento es sustituir las observaciones perdidas por la media de las observaciones disponibles. No añade ninguna información nueva y la media de todas las observaciones incluidas las observaciones perdidas es la misma. En el presente trabajo se realizará la comprobación.

Imputación múltiple (Múltiple Imputation, MI). En MI se generan valores imputados sobre la base de los datos existentes. En lugar de usar un solo valor para cada observación perdida (imputation simple o sencilla), el procedimiento IM reemplaza cada observación perdida por un conjunto de valores plausibles que representan la incertidumbre sobre el valor apropiado a imputar

Cuando hay observaciones perdidas, el software estadístico SPSS, automáticamente utiliza el método de supresión o el de datos completos, donde dichos datos faltantes se puede substituir por el **método de la media**. El software R también elimina automáticamente, las observaciones perdidas, y también se puede sustituir dichas observaciones faltantes por la media.

Los que promueven el uso de la imputación múltiple como el método más adecuado para reponer información omitida (Rubin, 1987), afirman que los procedimientos de MI generan buenos resultados, aun con porcentaje de omisión del 30,40 o 50 %. No obstante, es preciso señalar que cuando se trabaja con una encuesta probabilística, el tamaño de muestra garantiza cierta precisión para una tasa máxima de no respuesta, y en la medida de que la omisión supera el umbral establecido se pone en riesgo la confiabilidad estadística de las variables principales.

Por tanto se recomienda imputar datos en situaciones en que la omisión en una o más variables alcance porcentajes superiores al 20%.

En el contexto descrito, el objetivo principal de esta tesis es presentar las demostraciones de estimar las observaciones faltantes mediante la media aritmética y determinar su efecto en el análisis de componentes principales.

Para alcanzar el objetivo planteado, el presente trabajo se ha dividido en cuatro capítulos:

En el primer capítulo se presenta aspectos importantes sobre las observaciones faltantes en el contexto univariante y la estimación de parámetros poblacionales univariantes.

En el segundo capítulo se presenta aspectos importantes sobre las observaciones faltantes en el contexto multivariante y la estimación de los parámetros multivariantes.

En el tercer capítulo, se presenta el efecto de la estimación de observaciones faltantes en el análisis de componentes principales**.**

Finalmente en el cuarto capítulo, se valida los resultados de realizar el análisis de componentes principales con observaciones faltantes, las que han sido estimadas mediante la media aritmética de las observaciones conocidas de la respectiva variable, para lo cual se utilizó el software R.

CAPÍTULO I

MARCO TEÓRICO

OBSERVACIONES FALTANTES EN EL CONTEXTO UNIVARIANTE

1.1 INTRODUCCION

Los datos ausentes o faltantes son algo habitual en el análisis, la necesidad de centrarse en las razones de la ausencia de datos se desprende del hecho de que el investigador debe entender el proceso principal de la ausencia de datos. Cuando los procesos de ausencia de los datos son desconocidos, el investigador intenta identificar cualquier patrón en los datos faltantes o ausentes que caracterizaran dicho proceso.

Los datos longitudinales en los que cada sujeto o unidad experimental se mide u observa en ocasiones múltiples es probable que posean muchos datos perdidos, debido a que los sujetos no completan el estudio o salen antes de que el mismo finalice.

Se plantean cuestiones tales como: ¿están los datos faltantes distribuidos aleatoriamente entre las observaciones o se pueden identificar distintas pautas? ,¿En qué medida son relevantes?, se asume que está operando algún proceso de ausencia de datos y que algunos de los resultados estadísticos basados en esos datos podrían estar sesgados en la medida en que las variables incluidas en el análisis están incluidas por el proceso de pérdida de datos.

El impacto de los datos ausentes o faltantes es perjudicial no solo por sus potenciales sesgos sino también por su efecto en el tamaño de la muestra disponible para el análisis.

Cuando un conjunto de datos presenta pérdidas parciales en algunas variables y se desea estimar el vector de medias y la matriz de covariancias poblacional, un procedimiento habitual consiste en excluir las observaciones con datos faltantes, en cada una de las variables investigadas.

Todos los métodos generales de estimación de parámetros asumen que no existen datos faltantes en la muestra. Cuando existen observaciones perdidas en la muestra,

la solución más simple es eliminar aquellos individuos con observaciones incompletas y restringir el estudio a los individuos que presentan observaciones completas para todas las variables. Una consecuencia de este método es la reducción de individuos en la muestra respecto a la muestra planificada, lo que produce mayores sesgos en las estimaciones y mayor varianza muestral.

Por otro lado, la mayoría de los estudios relacionados con los métodos de imputación se centran en el uso de la media y su varianza y este basado en diseños muéstrales simples tales como el muestreo aleatorio simple.

En el trabajo se hará la justificación matemática del uso de la media aritmética en la estimación de observaciones faltantes o pérdidas y sus implicancias en la varianza, las correlaciones entre variables y en el análisis de componentes principales.

1.2 OBJETIVO GENERAL

Demostrar que la estimación de los datos faltantes mediante la media aritmética de los datos conocidos, no tiene efecto y/o modificación en el valor de la media aritmética de todos los datos, incluidos los datos faltantes previamente estimados por la media aritmética de los datos conocidos y sus consecuencias en el análisis de componentes Principales.

1.3 ANTECEDENTES DEL PROBLEMA

Según Juan A. Hernández Cabrera y Gustavo Ramírez Santana (1997), En las investigaciones del campo aplicado con técnicas multivariadas es muy frecuente encontrar matrices de datos con valores perdidos. Las estrategias más comúnmente utilizadas para reconducir este problema, utilizan los métodos listwise, pairwise y los de estimación de máxima verosimilitud. En este articulo se demuestra mediante las técnicas de simulación de Monte Carlo en el ámbito de los modelos estructurales, que independientemente del patrón de missing simulado (missing completamente aleatorio, monotónico o condicional) la estimación mediante el algoritmo de máxima verosimilitud EM arroja los mejores resultados, en cuanto a la precisión de la estimación de los parámetros de los modelos, disminución de los errores típicos, y la posibilidad de encontrar soluciones adecuadas y convergentes en aquellos patrones de missing donde las estrategias MCAR (listwise y pairwise) son imposibles de utilizar. A la luz da claramente mayor eficacia la estimación de máxima verosimilitud de las matrices de varianzas y covarianzas (utilizadas en todos las técnicas

estadísticas multivariadas), la conclusión de esta investigación recae en el hecho de recomendar la utilización de esta técnica para estimar la matriz de momentos siempre que el investigador se encuentre ante matrices de datos con valores perdidos independientemente de que el patrón sea MCAR o MAR.

Tal recomendación se sustenta en el hecho de que aunque la estrategia listwise es suficientemente eficiente en lo que a la estimación de los parámetros se refiere, en patrones missing completamente aleatorios y monotónicos, no lo es tanto en el estadístico de ajuste y en los errores típicos que son claramente más elevados que los de la muestra sin missing, lo que conducirá frecuentemente a la eliminación de parámetros "aparentemente no significativos" del modelo investigado.
Por otra parte, el número de soluciones convergentes y adecuadas con esta estrategia es claramente menor al conseguido con la estimación máximo verosímil(ML). Cuando el patrón de missing es MAR o el número de casos perdidos es muy elevado, puede producirse un sesgo en la estimación de los parámetros ya que la matriz muestral listwise no es una muestra aleatoria de la matriz de datos sin missing, o la imposibilidad de estimar el modelo dado que la matriz listwise contiene muy pocos casos. Tal y como hemos podido comprobar, en todas las ocasiones la estimación de máxima verosimilitud fue claramente superior a la realizada a partir de la matriz listwise, y esta estrategia fue imposible de utilizar cuando el patrón de missing era condicional. Hay que indicar, sin embargo, que la estimación ML en este patrón, aunque exitosa en las 500 muestras utilizadas, requirió de un número muy elevado de iteraciones (aproximadamente 200), dado que se utilizó como matriz de comienzo para iterar una matriz identidad de orden pxp (11 x 11).
En el caso de que se necesite disponer de los valores perdidos, y no solamente del vector de medias y de la matriz de varianzas y covarianzas, puede realizarse la triple imputación de los datos perdidos, una vez estimadas las matrices de momentos anteriores por ML, realizando posteriormente la ponderación de los casos por 1/3 para poder llevar acabo de esta forma los análisis multivariados clásicos con normalidad.
***Según Medina y Galván (2007),* un artículo publicado por la CEPAL.** Se suelen considerar aproximaciones para lidiar con datos perdidos (Little & Rubín, 2002)
Eliminar la información: en este caso se omite el registro de todo el análisis, con el consiguiente perjuicio de que podría haber diferencias sistemáticas entre usar o no la muestra completa, producir sesgos e incrementos en la dispersión. Cabe destacar que si la unidad de análisis es el país, eliminar el registro significaría eliminar el país, lo que podría llegar a ser inaceptable.
Alternativamente se puede ***eliminar la variable del análisis***. En este caso como regla empírica, se puede considerar que si una variable posee menos del 5% de datos perdidos respecto a todo el conjunto, no conviene eliminarla.

Hacer una ***imputación simple de los datos***, por ejemplo, a través del uso de promedios, medianas, modas, o mediante regresiones con la información disponible.
Imputación múltiple: en este caso se recurre a técnicas más sofisticadas como los algoritmos de Monte Carlo vía el uso de cadenas de Markov.

La principal ventaja de asignar datos perdidos es que con ello se reducen los sesgos y se realiza el análisis sobre la base de una cierta completitud en el conjunto de información. No obstante, la Incerteza que deviene de imputar datos debe quedar reflejada en la varianza de las estimaciones.

La Asignación simple de datos perdidos puede dar lugar la subestimación de la varianza. Al realizar imputación simple a los datos, siguiendo a Little y Rubín (2002), la asignación debe realizarse a partir de una distribución de probabilidades estimada a partir de la información disponible.

Según Schafer (1999) en los últimos años se han desarrollado una serie de métodos estadísticos para el análisis de datos completos.El algoritmo EM y sus extensiones, imputaciones multiples y el método de montecarlo en cadenas de markov, proporcionan un conjunto de métodos flexibles de inferencia estadística en problemas con datos perdidos.

1.4 ESTIMACIÓN DE PARAMETROS POBLACIONALES UNIVARIANTES

1.4.1 Aspectos generales

Se considera el caso de una única variable aleatoria, X , donde el interés es la estimación de los parámetros de locación μ y de dispersión σ^2.

En ese contexto, se toma una muestra aleatoria de tamaño $n+1$, $(x_1,...,x_n,x_{n+1})$, en la que una o más observaciones resultan faltantes. Sin pérdida de generalidad vamos a considerar que la observación que ocupa la posición n+1 ésima, es la observación faltante. Es decir , las observaciones $(x_1,...,x_n)$ son conocidas mientras que la x_{n+1} ésima es la observación faltante.

Sean las siguientes definiciones:

Definición 1.1.

La media muestral de la variable aleatoria, X , con "n" observaciones conocidas, $\bar{x}_n$ se define como:

$$\bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$$

Definición 1.2.

Varianza muestral sesgada con las n observaciones conocidas, S_n^2 se define como:

$$s_n^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}{n}$$

Definición 1.3.

Varianza muestral insesgada con las n observaciones conocidas, V_n^2 se define como:

$$V_n^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}{n-1}$$

Definición 1.4.

Estímese la $n+1$ ésima observación faltante, x_{n+1} mediante la media aritmética de las n observaciones conocidas. Es decir,

$$x_{n+1} = \bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (1)$$

Corolario 1.1. De la ecuación (1) se desprende que el cuadrado de la observación perdida es igual al cuadrado de la media de las observaciones conocidas,

$$\bar{x}_n \quad (x_{n+1})^2 = x_{n+1}^2 = (\bar{x}_{n+1})^2 = (\bar{x}_n)^2$$

Definición 1.5.

Media muestral de la variable X con $n+1$ observaciones, donde la $n+1$ ésima es la observación perdida o faltante, $\bar{x}_{n+1}$;se define como:

$$\bar{x}_{n+1} = \frac{\sum_{i=1}^{n+1} x_i}{n+1}$$

(2)

Definición 1.6

Varianza muestral sesgada con $n+1$ observaciones donde la $n+1$ ésima es la observación pérdida o faltante

s_{n+1}^2 : Se define como (3)

$$s_{n+1}^2 = \frac{\sum_{i=1}^{n+1}(x_i - \bar{x}_n)^2}{n+1}$$

Definición 1.7

Varianza muestral insesgada con $n+1$ observaciones donde la $n+1$ ésima es la observación perdida o faltante, V_{n+1}^2 ,se define como:

$$V_{n+1}^2 = \frac{\sum_{i=1}^{n+1}(x_i - \bar{x}_n)^2}{n}$$

Corolario 1.2

Para una nueva variable aleatoria y con $(y_1,.....y_n)$ observaciones donde la $n+1$ ésima es la observación perdida, donde

$\bar{y}_n = \frac{\sum_{i=1}^{n} y_i}{n}$ es la media muestral con las n observaciones conocidas, por lo que, la $n+1$ ésima observación, y_{n+1} , es la media muestral de las n observaciones conocidas.

$$y_{n+1} = \bar{y}_{n+1} = \frac{\sum_{i=1}^{n+1} y_i}{n+1}$$

Definición 1.8

La covarianza muestral de las variables X e Y para muestras de tamaño n, con todas las observaciones conocidas:

$$s_{XY(n)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{n-1}$$

Definición 1.9 La covarianza muestral de las variables X e Y para muestras de tamaño n+1 , donde las observaciones perdidas, ubicadas en cada caso en cualquier posición, sin pérdida de generalidad, han sido estimadas mediante las respectivas medias de la muestra para cada una de las variables separadamente. Es decir, $x_{n+1} = \bar{x}_n \quad y_{n+1} = \bar{y}_{n+1}$

Entonces,

$$s_{XY(n+1)} = \frac{\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})(y_i - \bar{y}_{n+1})}{n}$$

Definición 1.10. El coeficiente de correlación muestral entre las variables X e Y, para muestras de tamaño n , $r_{(n)}$, se define como:

$$r_{(n)} = \frac{s_{XY(n)}}{s_{X(n)}\ s_{Y(n)}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}. \qquad (4)$$

Definición 1.11. El coeficiente de correlación muestral entre las variables X e Y con $n+1$ observaciones, r_{n+1}, contendiendo las observaciones perdidas es :

$$r_{(n+1)} = \frac{s_{XY(n+1)}}{s_{X(n+1)}\ s_{Y(n+1)}}$$

1.4.2 Resultados importantes

Teorema 1.1 Obtención de la media muestral con datos faltantes

Sea $x_1, x_2, \ldots, x_n, x_{n+1}$ una muestra aleatoria de tamaño $n+1$, donde n observaciones , $x_1, x_2, \ldots, x_n$, son conocidas con media aritmética $\bar{x}_n$

y la x_{n+1} ésima es la observación perdida. Entonces, si $x_{n+1} = \bar{x}_n = \dfrac{\sum_{i=1}^{n} x_i}{n}$,

la media muestral de todas las observaciones incluida la observación perdida, es igual a la media de las n observaciones conocidas. Es decir :

$$\bar{x}_{n+1} = \bar{x}_n .$$

(5)

Demostración:

La media muestral para la observación perdida n+1 ésima se obtiene de la siguiente manera:

$$\begin{aligned}
\bar{x}_{n+1} &= \sum_{i=1}^{n+1} \frac{x_i}{n+1} \\
&= \frac{\sum_{i=1}^{n} x_i}{n+1} + \frac{x_{n+1}}{n+1} \\
&= \frac{n\bar{x}_n}{n+1} + \frac{\bar{x}_n}{n+1} \\
&= \frac{(n+1)\bar{x}_n}{(n+1)} \\
&= \bar{x}_n
\end{aligned}$$

Es decir

$$\bar{x}_{n+1} = \bar{x}_n$$

Teorema 1.2.- Sea $x_1, x_2, \ldots, x_n, x_{n+1}$ una muestra aleatoria de tamaño n+1 , donde n observaciones , $x_1, x_2, \ldots, x_n$, son conocidas con media aritmética $\bar{x}_n$

y la x_{n+1} ésima es la observación perdida.

Si

$$x_{n+1} = \bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$$, entonces

$$\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})^2 = \sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2 .$$

(6)

Demostración:

$$\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1}) =$$

$$\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1}) = \sum_{i=1}^{n+1}(x_i^2 - 2x_i\bar{x}_{n+1} + (\bar{x}_{n+1})^2)$$

$$= \sum_{i=1}^{n+1}(x_i^2 - 2(n+1)\bar{x}_{n+1}\bar{x}_{n+1} + (n+1)\bar{x}_{n+1}^2)$$

$$= \sum_{i=1}^{n+1}(x_i^2) - 2(n+1)\bar{x}_{n+1}\bar{x}_{n+1} + (n+1)\bar{x}_{n+1}^2)$$

$$= \sum_{i=1}^{n+1}(x_i^2) - 2(n)\bar{x}_{n+1}^2 - 2\bar{x}_{n+1}^2 + n\bar{x}_{n+1}^2 + \bar{x}_{n+1}^2$$

$$= \sum_{i=1}^{n} x_i^2 + \bar{x}_n^2 - 2(n)\bar{x}_{n+1}^2 - 2\bar{x}_{n+1}^2 + n\bar{x}_{n+1}^2 + \bar{x}_{n+1}^2$$

$$= \sum_{i=1}^{n} x_i^2 + \bar{x}_n^2 - n\bar{x}_{n+1}^2 - \bar{x}_{n+1}^2$$

$$= \sum_{i=1}^{n} x_i^2 + \bar{x}_n^2 - n\bar{x}_{n+1}^2 - \bar{x}_n^2$$

$$= \sum_{i=1}^{n} x_i^2 \quad - n\bar{x}_n^2$$

$$= \sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2$$

Entonces,

$$\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})^2 = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

Por el Corolario 1. 2 Para la muestra $y_1, y_2,\ldots, y_n, y_{n+1}$

una muestra de tamaño $n+1$, $\bar{y}_{n+1} = \bar{y}_n$ y $\sum_{i=1}^{n+1}(y_i - \bar{y}_{n+1})^2 = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2$

Teorema 1.3 Obtención de la varianza muestral sesgada con datos faltantes

Sea $x_1, x_2,\ldots, x_n, x_{n+1}$ una muestra aleatoria de tamaño $n+1$, donde n observaciones, $x_1, x_2,\ldots, x_n$, son conocidas con media aritmética $\bar{x}_n$

y la x_{n+1} ésima es la observación perdida $x_{n+1} = \bar{x}_n = \dfrac{\sum_{i=1}^{n} x_i}{n}$ .Entonces,

$S_{n+1}^2 = \dfrac{n}{n+1} S_n^2$, es la varianza muestral sesgada.

(7)

Demostración:

Sea

$$s_{n+1}^2 = \frac{\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})^2}{n+1}$$

Según definición (1.6), desarrollando el numerador y reemplazando el resultado de la ecuación (2)

$$\sum_{i=1}^{n+1}\left(x_i-\bar{x}_{n+1}\right)^2=\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2$$

$$s_{n+1}^2=\frac{\sum_{i=1}^{n+1}\left(x_i-\bar{x}_n\right)^2}{n+1}=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2}{n+1}$$

$$s_{n+1}^2=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2}{n}\frac{n}{n+1}$$

$$S_{n+1}^2=\frac{n}{n+1}S_n^2$$

Se ha probado que la varianza muestral sesgada con el dato faltante, es igual a la varianza muestral sesgada con n observaciones multiplicada por el factor $\frac{n}{n+1}$.
es decir

$$s_{n+1}^2=\frac{n}{n+1}s_n^2$$

Teorema 1.4 Obtención de la varianza muestral insesgada con datos faltantes

Sea $x_1,x_2,...,x_n,x_{n+1}$ una muestra aleatoria de tamaño $n+1$, donde n observaciones, $x_1,x_2,...,x_n$, son conocidas con media aritmética $\bar{x}_n$

y la x_{n+1} ésima es la observación perdida.

$x_{n+1}=\bar{x}_n=\frac{\sum_{i=1}^{n}x_i}{n}$, entonces,

la varianza muestral insesgada con el dato faltante es $V_{n+1}^2=\frac{n-1}{n}V_n^2$.

(8)

Demostración:

Sea $V_{n+1}^2 = \dfrac{\sum_{i=1}^{n+1}\left(x_i - \bar{x}_{n+1}\right)^2}{n}$

Desarrollando

$$\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})^2 = \sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2$$

$$V_{n+1}^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2}{n} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2}{n}\left(\frac{n-1}{n-1}\right) =$$

$$= \frac{n-1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}{n-1}$$

$$= \frac{n-1}{n}V_n^2$$

Entonces

$V_n^2 = \dfrac{n-1}{n}V_n^2$

Es decir la varianza muestral insesgada con el dato faltante es igual a la varianza muestral insesgada para las n observaciones multiplicado por el factor $\dfrac{n-1}{n}$

Corolario 1. 3

Sea $x_1, x_2, ..., x_n, x_{n+1}$ una muestra aleatoria de tamaño $n+1$, donde n observaciones, $x_1, x_2, ..., x_n$, son conocidas con media aritmética $\bar{x}_n$

y la x_{n+1} ésima es la observación perdida. Entonces, si $x_{n+1} = \bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$, entonces,

$$V_{n+1}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}{n} = s_n^2$$

Demostración

Se tiene que

$$V_{n+1}^2 = \frac{n-1}{n} V_n^2 =$$

$$= \frac{n-1}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}{n-1} =$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}{n} = S_n^2$$

Lo que queda demostrado que

$$V_{n+1}^2 = S_n^2$$

Es decir la varianza muestral insesgada con el dato faltante es igual a la varianza muestral sesgada para n observaciones.

1.5 Estimación Parámetros Poblacionales Bivariantes

En este contexto, será el coeficiente de correlación poblacional el parámetro poblacional.

Teorema 1.5 Obtención del Coeficiente de Correlación lineal de Pearson con datos faltantes

Sea $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ el parámetro poblacional coeficiente de correlación lineal de Pearson entre las variables X e Y, donde $\sigma_{XY}, \sigma_X, \sigma_Y$ son la covarianza entre las dos variables y las desviaciones estándar de cada una de ellas.

Sean $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), (x_{n+1}, y_{n+1})$ una muestra aleatoria de tamaño n+1, donde sin pérdida de generalidad puede considerarse $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ las observaciones conocidas y mientras que (x_{n+1}, y_{n+1}) es la observación perdida·

Si los datos faltantes $x_{n+1,} y_{n+1}$ se estiman con la media de las observaciones conocidas en cada caso, es decir para cada variable, $x_{n+1} = \bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$, $y_{n+1} = \bar{y}_n = \frac{\sum_{i=1}^{n} y_i}{n}$, entonces, el coeficiente de correlación lineal muestral con todas las observaciones donde las observaciones perdidas se han estimado mediante el promedio de las observaciones conocidas, $r_{(n+1)}$, es igual al coeficiente de correlación lineal con solo las observaciones conocidas, $r_{(n)}$· Es decir,

$$r_{(n+1)} = r_{(n)} . \qquad (9)$$

donde $r_{(n)}$ ha sido definida en la ecuación (4).

Demostración:

Por definición:

$$r_{(n+1)} = \frac{\frac{\sum_{i=1}^{n+1} (x_i - \bar{x}_{n+1})(y_i - \bar{y}_{n+1})}{n}}{\sqrt{\frac{\sum_{i=1}^{n+1} (x_i - \bar{x}_{n+1})^2}{n} \frac{\sum_{i=1}^{n+1} (y_i - \bar{y}_{n+1})^2}{n}}}$$

$$r_{n+1} = \frac{\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})(y_i - \bar{y}_{n+1})}{\sqrt{\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})^2 \sum_{i=1}^{n+1}(y_i - \bar{y}_{n+1})^2}}$$

Desarrollando el numerador se tiene que:

$$= \sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})(y_i - \bar{y}_{n+1}) =$$

$$= \sum_{i=1}^{n+1}(x_i y_i - \bar{x}_{n+1}\bar{y}_{n+1} - x_i \bar{y}_{n+1} + \bar{x}_{n+1}\bar{y}_{n+1})$$

$$= \sum_{i=1}^{n+1} x_i y_i - (n+1)\bar{x}_{n+1}\bar{y}_{n+1} - (n+1)\bar{x}_{n+1}\bar{y}_{n+1} + (n+1)\bar{x}_{n+1}\bar{y}_{n+1}$$

$$= \sum_{i=1}^{n+1} x_i y_i - n\bar{x}_{n+1}\bar{y}_{n+1} - \bar{x}_{n+1}\bar{y}_{n+1} - n\bar{x}_{n+1}\bar{y}_{n+1} - \bar{x}_{n+1}\bar{y}_{n+1} + n\bar{x}_{n+1}\bar{y}_{n+1} + \bar{x}_{n+1}\bar{y}_{n+1}$$

$$= \sum_{i=1}^{n+1} x_i y_i - \bar{x}_{n+1}\bar{y}_{n+1} - n\bar{x}_{n+1}\bar{y}_{n+1}$$

$$= \sum_{i=1}^{n} x_i y_i + x_{n+1} y_{n+1} - \bar{x}_{n+1}\bar{y}_{n+1} - n\bar{x}_{n+1}\bar{y}_{n+1}$$

$$= \sum_{i=1}^{n} x_i y_i + \bar{x}_n \bar{y}_n - \bar{x}_n \bar{y}_n - n\bar{x}_n \bar{y}_n$$

$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}_n \bar{y}_n$$

$$= \sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)$$

$$= \sum_{i=1}^{n}(x_i - \bar{x}_{n+1})(y_i - \bar{y}_{n+1})$$

$$= \sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)$$

Con respecto al denominador, por (6) se sabe que:

$\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})^2 = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ y respectivamente

$\sum_{i=1}^{n+1}(y_i - \bar{y}_{n+1})^2 = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2$.

Luego:

$$r_{n+1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}$$

$$r_{n+1} = r_n$$

Se ha probado que el coeficiente de correlación lineal es invariante al número de observaciones.

CAPITULO II

OBSERVACIONES PERDIDAS EN EL CONTEXTO MULTIVARIANTE

2.1 Aspectos generales

Los métodos estadísticos multivariantes y el análisis multivariante son herramientas estadísticas que estudian el comportamiento de dos o más variables al mismo tiempo.

Se generalizará los resultados del capítulo anterior en la estimación de observaciones faltantes al caso multivariante, es decir, en la estimación de los parámetros multivariantes, vector de medias poblacional, matriz de varianzas y covarianzas poblacional y matriz de correlaciones poblacional.

Sean $\vec{X} = (X_1, ..., X_p)^T$ Es un vector aleatorio de dimensión p con

$E(\vec{X}) = \vec{\mu}$ y matriz de covarianzas de Σ,

donde

$$\vec{\mu} = E\begin{bmatrix} X_1 \\ X_2 \\ ... \\ ... \\ X_P \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ ... \\ ... \\ u_p \end{bmatrix} \quad \text{y} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & ... & ... & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & ... & ... & \sigma_{2p} \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ \sigma_{p1} & \sigma_{p2} & & & \sigma_{pp} \end{bmatrix}$$

son el vector de medias y la matriz de covarianzas poblacionales respectivamente.

Dado que cada observación perdida, para cada variable se puede estimar separadamente usando la misma metodología descrita en el capítulo anterior, con fines de estimación de los parámetros descritos, y sin pérdida de generalidad, considérese que $\vec{x}_1, ..., \vec{x}_n, \vec{x}_{n+1}$ es una muestra aleatoria de tamaño $n+1$ desde la población multivariante, donde cada $\vec{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^{\mathrm{T}}$ $i = 1,2,.....,p$,

contiene las observaciones conocidas y $\vec{x}_{n+1}$ es la observación multivariante perdida.

Definición 2.1 $\mathbf{X}_{(n)}$, es la matriz de datos multivariante

$$\mathbf{X}_{(n)} = \left(\vec{x}_1,....,\vec{x}_n\right)^T = \begin{pmatrix} x_{11} & x_{12} & . & . & x_{1p} \\ x_{21} & x_{22} & . & . & x_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{n1} & . & . & . & x_{np} \end{pmatrix}$$ donde x_{ij} es la medida de la

j -ésima variable en el i -ésimo individuo, $i = 1,2,.....,n;\quad j = 1,...p$.

Definición 2.2

$\vec{\bar{x}}_{(n)}$: es el vector de la media muestral para n observaciones o cuando el tamaño de la muestra es n , con todos los datos conocidos.

$$\vec{\bar{x}}_{(n)} = \begin{bmatrix} \bar{x}_{1(n)} \\ \bar{x}_{2(n)} \\ . \\ . \\ . \\ \bar{x}_{p(n)} \end{bmatrix} = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i2} \\ \cdots \\ \cdots \\ \sum_{i=1}^{n} x_{ip} \end{bmatrix} = \frac{1}{n}\begin{pmatrix} x_{11} & . & . & . & x_{1p} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{n1} & . & . & . & x_{np} \end{pmatrix}^T$$

$$= \frac{1}{n}\mathbf{X}_{(n)}^T$$

Definición 2.3 Definase como, $X_{(n+1)}$ es la matriz de datos con $n+1$ observaciones, donde sin pérdida de generalidad, la $n+1$ é-sima observación multivariante contiene datos faltantes.

$$X_{(n+1)} = \begin{bmatrix} x_{11} & x_{12} & \dots & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & \dots & x_{np} \\ x_{n+1,1} & x_{n+1,2} & \dots & \dots & x_{n+1,p} \end{bmatrix}$$

Definición 2.4 *Sea* $\vec{x}_{(n+1)}$*: vector de observaciones faltantes o la* $n+1$ *ésima* observación multivariante con datos faltantes.Se define como el vector de medias de las n observaciones y vamos a considerar que ocupan las primeras n observaciones conocidas. Usando los resultados del capítulo 1.

$$\vec{x}_{(n+1)} = (x_{n+1,1}, x_{n+1,2}, \dots\dots x_{n+1,p})$$

donde

$x_{n+1,1}$: *es la observación perdida en la primera variable,* $x_{n+1,1} = \bar{x}_{1(n)}$

$x_{n+1,j}$: es la observación perdida en la j-ésima variable, $x_{n+1,2} = \bar{x}_{2(n)}$

$x_{n+1,p}$: es la observación perdida en la p-ésima variable, $x_{n+1,p} = \bar{x}_{p(n)}$

$\bar{x}_{1(n)}$: es la media muestral de la variable con las primeras n observaciones conocidas.

$\bar{x}_{2(n)}$: es la media muestral de la variable con las primeras n observaciones conocidas

$\bar{x}_{j(n)}$: es la media muestral de la j-ésima variable con n observaciones conocidas.

$\bar{x}_{p(n)}$: es la media muestral de la p-ésima variable con n observaciones conocidas.

$x_{n+1,1} = \bar{x}_{1(n)}$: es la observación perdida en la j é-sima variable resulta siendo la media de las n observaciones conocidas en la j e´sima variable y asi sucesivamente hasta

$x_{n+1,p} = \bar{x}_{p(n)}$: es la observación perdida en la p-é-sima variable resulta siendo la media de las n observaciones conocidas en la p-é-sima variable.

Definición 2.5

$S_{(n)}$: Matriz de covarianzas muestral sesgada para la muestra con las n observaciones conocidas, donde

$$S_n = \frac{1}{n}\left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)^T \left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)$$

$$= \frac{1}{n} X_{(n)}^T X_{(n)} - \vec{\bar{X}}_{(n)} \vec{\bar{X}}_{(n)}^T .$$

(10)

$$. \; S_{(n)} = \begin{bmatrix} s_{11} & s_{12} & \dots & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & \dots & s_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ s_{p1} & s_{p2} & \dots & \dots & s_{pp} \end{bmatrix}$$

$$S_{(n)} = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n}\left(x_{i1} - \bar{x}_{1(n)}\right)^2 & \sum_{i=1}^{n}\left(x_{i1} - \bar{x}_{1(n)}\right)\left(x_{i2} - \bar{x}_{2(n)}\right) & \dots & \dots & \sum_{i=1}^{n}\left(x_{i1} - \bar{x}_{1(n)}\right)\left(x_{ip} - \bar{x}_{p(n)}\right) \\ \sum_{i=1}^{n}\left(x_{i2} - \bar{x}_{2(n)}\right)\left(x_{i1} - \bar{x}_{1(n)}\right) & \sum_{i=1}^{n}\left(x_{i2} - \bar{x}_{2(n)}\right)^2 & \dots & \dots & \sum_{i=1}^{n}\left(x_{i2} - \bar{x}_{2(n)}\right)\left(x_{ip} - \bar{x}_{p(n)}\right) \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i=1}^{n}\left(x_{ip} - \bar{x}_{p(n)}\right)\left(x_{i1} - \bar{x}_{1(n)}\right) & \sum_{i=1}^{n}\left(x_{ip} - \bar{x}_{p(n)}\right)\left(x_{i2} - \bar{x}_{2(n)}\right) & \dots & \dots & \sum_{i=1}^{n}\left(x_{ip} - \bar{x}_{p(n)}\right)^2 \end{bmatrix}$$

Luego

$$s_{(n)} = \frac{1}{n}\left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)^T\left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)$$

$$= \frac{1}{n}(X^T_{(n)} - \vec{\bar{X}}^T_{(n)})(X_{(n)} - \vec{\bar{X}}^T{}_{(n)})$$

$$= \frac{1}{n}(X^T_{(n)}X_{(n)} - \vec{\bar{X}}^T_{(n)}X_{(n)} - X^T_{(n)}\vec{\bar{X}}_{(n)} + \vec{\bar{X}}^T_{(n)}\vec{\bar{X}}_{(n)})$$

$$= \frac{1}{n}X^T_{(n)}X_{(n)} - \frac{1}{n}\vec{\bar{X}}^T_{(n)}X_{(n)} - \frac{1}{n}X^T_{(n)}\vec{\bar{X}}_{(n)} + \frac{1}{n}\vec{\bar{X}}^T_{(n)}\vec{\bar{X}}_{(n)}$$

$$= \frac{1}{n}X^T_{(n)}X_{(n)} - \frac{1}{n}\vec{\bar{X}}^T_{(n)}\vec{\bar{X}}_{(n)} - \frac{1}{n}X^T_{(n)}\vec{\bar{X}}_{(n)} + \frac{1}{n}\vec{\bar{X}}^T_{(n)}\vec{\bar{X}}_{(n)}$$

$$= \frac{1}{n}X^T_{(n)}X_{(n)} - \frac{1}{n}X^T_{(n)}\vec{\bar{X}}_{(n)}$$

$$= \frac{1}{n}X^T_{(n)}X_{(n)} - \frac{1}{n}\left(n\vec{\bar{X}}_{(n)}\vec{\bar{X}}^T_{(n)}\right)$$

$$= \frac{1}{n}X^T_{(n)}X_{(n)} - \vec{\bar{X}}_{(n)}\vec{\bar{X}}^T_{(n)}$$

$$= \frac{1}{n}X^T_{(n)}X_{(n)} - \vec{\bar{X}}_{(n)}\vec{\bar{X}}^T_{(n)}$$

Definición 2.6

$V_{(n)}$ es la matriz de varianzas y covarianzas muestral insesgada de las p variables con n observaciones conocidas.

$$V_{(n)} = \begin{bmatrix} v_{11} & v_{12} & \cdots & \cdots & v_{1p} \\ v_{21} & v_{22} & \cdots & \cdots & v_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ v_{p1} & v_{p2} & \cdots & \cdots & v_{pp} \end{bmatrix}$$

Donde $V_n = \frac{1}{n-1}\left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)^T\left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)$

$$= \frac{1}{n-1}X^T_{(n)}X_{(n)} - \frac{n}{n-1}\vec{\bar{X}}_{(n)}\vec{\bar{X}}^T_{(n)}.$$

2.2 Estimaciones con datos faltantes

Teorema 2.1 Obtención del Vector de Medias con datos faltantes

Sin pérdida de generalidad, considérese que $\vec{x}_1,...,\vec{x}_n,\vec{x}_{n+1}$ una muestra aleatoria de tamaño $n+1$ desde la población multivariante, donde cada $\vec{x}_i = (x_{i1}, x_{i2},....., x_{ip})'$ $i = 1,2,.....,n$, contiene las observaciones conocidas y $\vec{x}_{(n+1)}$ es la observación multivariante perdida.

$X_{(n+1)}$ es la matriz de datos multivariante para la muestra de tamaño $n+1$, sin pérdida de generalidad, considerese la $n+1$ é-sima observación multivariante con datos faltantes.

$\vec{x}_{(n+1)}$:vector con datos faltantes estimados con las medias de las observaciones conocidas.

$$\vec{x}_{(n+1)} = (x_{n+1,1}, x_{n+1,2},....., x_{n+1,p})$$

$$= (\bar{x}_{1(n)}, \bar{x}_{2(n)},....., \bar{x}_{p(n)})$$

Entonces, el vector de media muestral con las n+1 observaciones es,

$$\vec{\bar{x}}_{(n+1)} = \vec{\bar{x}}_{(n)} = \begin{bmatrix} \bar{x}_{1(n)} \\ \bar{x}_{2(n)} \\ \cdot \\ \cdot \\ \cdot \\ \bar{x}_{p(n)} \end{bmatrix}^T$$

Demostración:

Partiendo de que

$$\vec{\bar{x}}_{(n+1)} = \begin{bmatrix} \bar{x}_{1(n+1)} \\ \bar{x}_{2(n+1)} \\ \cdot \\ \cdot \\ \cdot \\ \bar{x}_{p(n+1)} \end{bmatrix}$$

donde

$\vec{\bar{x}}_{(n+1)}$ es la media de la i-ésima variable con $n+1$ observaciones.

$$\vec{\bar{X}}_{(n+1)} = \frac{1}{n+1}\begin{bmatrix} \sum_{i=1}^{n+1} x_{i1} \\ \sum_{i=1}^{n+1} x_{i2} \\ \cdots \\ \cdots \\ \sum_{i=1}^{n+1} x_{ip} \end{bmatrix}$$

$$= \frac{1}{n+1}\begin{bmatrix} \sum_{i=1}^{n} x_{i1} + x_{n+1,1} \\ \sum_{i=1}^{n} x_{i2} + x_{n+1,2} \\ \cdots \\ \cdots \\ \sum_{i=1}^{n} x_{ip} + x_{n+1,p} \end{bmatrix}$$

Desarrollando cada término del vector, se tiene que :

$$\sum_{i=1}^{n} x_{i1} + x_{n+1,1} = n\bar{x}_{1(n)} + \bar{x}_{1(n)} = (n+1)\bar{x}_{1(n)}$$

$$\sum_{i=1}^{n} x_{ip} + x_{n+1,p} = n\bar{x}_{p(n)} + \bar{x}_{p(n)} = (n+1)\bar{x}_{p(n)}.$$

Entonces en forma matricial se tiene que

$$= \frac{1}{n+1}\begin{bmatrix} n\bar{x}_{1(n)} + \bar{x}_{1(n)} \\ n\bar{x}_{2(n)} + \bar{x}_{2(n)} \\ \cdots \\ \cdots \\ n\bar{x}_{p(n)} + \bar{x}_{p(n)} \end{bmatrix}$$

$$= \frac{1}{n+1}\begin{bmatrix} (n+1)\bar{x}_{1(n)} \\ (n+1)\bar{x}_{2(n)} \\ \cdots \\ \cdots \\ (n+1)\bar{x}_{p(n)} \end{bmatrix} = \frac{1}{n} X^{T}_{(n)} = \vec{\bar{X}}_{(n)}$$

Luego, se ha demostrado que

$$\vec{\bar{x}}_{(n+1)} = \vec{\bar{x}}_{(n)}$$

El vector de medias muestral no se afecta por la estimación de observaciones faltantes mediante la media de las observaciones conocidas.

Para la obtención tanto de la matriz de covarianzas muestral sesgada como insesgada, con observaciones perdidas, se hace necesario saber qué sucede con el numerador de ambas expresiones, es decir con las sumas de cuadrados correspondientes.

Definición 2.8. Se define la matriz de medias donde cada observación, de la matriz de datos conocida, ha sido reemplazada por su respectiva media. Es decir,

$$\vec{\bar{X}}_{(n)} = \begin{pmatrix} \bar{x}_{1(n)} & \bar{x}_{2(n)} & \cdots & \cdots & \bar{x}_{p(n)} \\ \bar{x}_{1(n)} & \bar{x}_{2(n)} & \cdots & \cdots & \bar{x}_{p(n)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \bar{x}_{1(n)} & \bar{x}_{2(n)} & \cdots & \cdots & \bar{x}_{p(n)} \end{pmatrix}$$

Matriz de orden $n_x n$ que contiene las medias de las p variables para muestras de tamaño n con todas las observaciones conocidas.

Definición 2.9 La matriz que contiene las sumas de cuadrados y productos cruzados con elementos necesarios para la obtención de las matrices de covarianzas muestrales, para las n observaciones conocidas es la siguiente:

$$\begin{bmatrix} \sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)^2 & \sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right) & \cdots & \cdots & \sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{ip}-\bar{x}_{p(n)}\right) \\ \sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n)}\right)\left(x_{i1}-\bar{x}_{1(n)}\right) & \sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n)}\right)^2 & \cdots & \cdots & \sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n)}\right)\left(x_{ip}-\bar{x}_{p(n)}\right) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n}\left(x_{ip}-\bar{x}_{p(n)}\right)\left(x_{i1}-\bar{x}_{1(n)}\right) & \sum_{i=1}^{n}\left(x_{ip}-\bar{x}_{p(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right) & \cdots & \cdots & \sum_{i=1}^{n}\left(x_{ip}-\bar{x}_{p(n)}\right)^2 \end{bmatrix}$$

Definición 2.10 La matriz $\vec{\bar{X}}_{(n+1)}$, de orden $(n+1)_x(n+1)$, cuyas columnas son las medias de cada una de las variables, toma la siguiente forma:

$$\vec{\bar{X}}_{(n+1)} = \begin{pmatrix} \bar{x}_{1(n+1)} & \bar{x}_{2(n+1)} & \cdots & \cdots & \bar{x}_{p(n+1)} \\ \bar{x}_{1(n+1)} & \bar{x}_{2(n+1)} & \cdots & \cdots & \bar{x}_{p(n+1)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \bar{x}_{1(n+1)} & \bar{x}_{2(n+1)} & \cdots & \cdots & \bar{x}_{p(n+1)} \end{pmatrix} = \begin{pmatrix} \bar{x}_{1(n)} & \bar{x}_{2(n)} & \cdots & \cdots & \bar{x}_{p(n)} \\ \bar{x}_{1(n)} & \bar{x}_{2(n)} & \cdots & \cdots & \bar{x}_{p(n)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \bar{x}_{1(n)} & \bar{x}_{2(n)} & \cdots & \cdots & \bar{x}_{p(n)} \end{pmatrix}$$

Teorema 2.2

Obtención de la matriz de covarianzas muestral sesgada con datos faltantes

Si la $n+1$ésima observación perdida en la $j$ésima variable, $x_{j(n+1)}$ se estima en la medida de las n observaciones conocidas de la misma variable , $\bar{x}_{j(n)}$, es decir si $x_{j,n+1} = \bar{x}_{j(n)}$, $j = 1.....p$, entonces la variable muestral sesgada con $n+1$ observaciones perdidas , $S_{(n+1)}$ está afectada por el factor $\frac{n}{n+1}$, con respecto a la matriz de covarianzas muestral con todas las observaciones conocidas, $\mathrm{S}_{(n)}$ es decir:

$$\mathrm{S}_{(n+1)} = \frac{n}{n+1}\mathrm{S}_{(n)} \qquad (12)$$

Demostración:

$\vec{x}_{n+1} = \left(x_{n+1,1}, x_{n+1,2}, ..., x_{n+1,p}\right)$ es un vector de dimensión $1\mathrm{x}(p)$,definido por

$\vec{x}_{n+1} = \left(\bar{x}_{1(n)}, \bar{x}_{2(n)}, ..., \bar{x}_{p(n)}\right)$, es decir: $\vec{x}_{n+1} = \vec{\bar{x}}^{T}_{(n)}$.

Se define la matriz de medias con "$n+1$" observaciones

$$\vec{\overline{X}}_{(n+1)} = \begin{pmatrix} \overline{x}_{1(n+1)} & \overline{x}_{2(n+1)} & \cdots & \cdots & \overline{x}_{p(n+1)} \\ \overline{x}_{1(n+1)} & \overline{x}_{2(n+1)} & \cdots & \cdots & \overline{x}_{p(n+1)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \overline{x}_{1(n+1)} & \overline{x}_{2(n+1)} & \cdots & \cdots & \overline{x}_{p(n+1)} \end{pmatrix}$$

Consideremos la matriz de sumas de cuadrados y productos cruzados con $n+1$ observaciones:

$$\begin{bmatrix} \sum_{i=1}^{n+1}\left(x_{i1}-\overline{x}_{1(n)}\right)^2 & \sum_{i=1}^{n+1}\left(x_{i1}-\overline{x}_{1(n)}\right)\left(x_{i2}-\overline{x}_{2(n)}\right) & \cdots & \cdots & \sum_{i=1}^{n+1}\left(x_{i1}-\overline{x}_{1(n)}\right)\left(x_{ip}-\overline{x}_{p(n)}\right) \\ \sum_{i=1}^{n+1}\left(x_{i2}-\overline{x}_{2(n)}\right)\left(x_{i1}-\overline{x}_{1(n)}\right) & \sum_{i=1}^{n+1}\left(x_{i2}-\overline{x}_{2(n)}\right)^2 & \cdots & \cdots & \sum_{i=1}^{n+1}\left(x_{i2}-\overline{x}_{2(n)}\right)\left(x_{ip}-\overline{x}_{p(n)}\right) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n+1}\left(x_{ip}-\overline{x}_{p(n)}\right)\left(x_{i1}-\overline{x}_{1(n)}\right) & \sum_{i=1}^{n+1}\left(x_{ip}-\overline{x}_{p(n)}\right)\left(x_{i2}-\overline{x}_{2(n)}\right) & \cdots & \cdots & \sum_{i=1}^{n+1}\left(x_{ip}-\overline{x}_{p(n)}\right)^2 \end{bmatrix}$$

Para las $n+1$ observaciones se definen las varianzas de cada variable, se tiene que

$$s_{jj} = \frac{\sum_{i=1}^{n+1}\left(x_{ij}-x_{j(n)}\right)^2}{n+1} \quad j=1...p$$

A continuación estamos desarrollando la parte del numerador de cada varianza , con $"n+1"$observaciones.

$$\sum_{i=1}^{n+1}\left(x_{i1}-\overline{x}_{1(n)}\right)^2 = \sum_{i=1}^{n}\left(x_{i1}-\overline{x}_{1(n)}\right)^2 + \left(x_{n+1,1}-\overline{x}_{1(n)}\right)^2$$

$$\sum_{i=1}^{n+1}\left(x_{i1}-\overline{x}_{1(n)}\right)^2 = \sum_{i=1}^{n}\left(x_{i1}-\overline{x}_{1(n)}\right)^2 + \left(\overline{x}_{1(n)}-\overline{x}_{1(n)}\right)^2$$

$$\sum_{i=1}^{n+1}\left(x_{i1}-\overline{x}_{1(n)}\right)^2 = \sum_{i=1}^{n}\left(x_{i1}-\overline{x}_{1(n)}\right)^2 .$$

Para la variable "j" e-sima se tiene que la suma de cuadrados

$$\sum_{i=1}^{n+1}\left(x_{ij}-\bar{x}_{j(n)}\right)^2=\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)^2+\left(x_{n+1,j}-\bar{x}_{j(n)}\right)^2$$

$$=\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)^2+\left(\bar{x}_{j(n)}-\bar{x}_{j(n)}\right)^2$$

$$=\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)^2 \qquad j=1,..,p$$

O sea: $$\sum_{i=1}^{n+1}\left(x_{ij}-\bar{x}_{j(n)}\right)^2=\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)^2 \qquad j=1,..,p$$

De la misma forma para la covarianza de la primera con la segunda variable, la parte del numerador se obtiene de la siguiente manera:

$$\sum_{i=1}^{n+1}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right)=\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right)+\left(x_{n+1,1}-\bar{x}_{1(n)}\right)\left(x_{n+1,2}-\bar{x}_{2(n)}\right)$$

$$=\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right)+\left(\bar{x}_{1(n)}-\bar{x}_{1(n)}\right)\left(\bar{x}_{2(n)}-\bar{x}_{2(n)}\right)$$

$$=\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right)$$

Es decir: $$\sum_{i=1}^{n+1}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right)=\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right)$$

Para la covarianza de la j-e-sima variable y k-é-sima variable la parte del numerador se demuestra de la siguiente manera:

$$\sum_{i=1}^{n+1}\left(x_{ij}-\bar{x}_{j(n)}\right)\left(x_{ik}-\bar{x}_{k(n)}\right)=\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)\left(x_{ik}-\bar{x}_{k(n)}\right)+\left(x_{n+1,j}-\bar{x}_{j(n)}\right)\left(x_{n+1,k}-\bar{x}_{k(n)}\right)$$

$$=\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)\left(x_{ik}-\bar{x}_{k(n)}\right)+\left(\bar{x}_{j(n)}-\bar{x}_{j(n)}\right)\left(\bar{x}_{k(n)}-\bar{x}_{k(n)}\right)$$

$$= \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_{j(n)}\right)\left(x_{ik} - \bar{x}_{k(n)}\right)$$

$$\sum_{i=1}^{n+1} \left(x_{ij} - \bar{x}_{j(n)}\right)\left(x_{ik} - \bar{x}_{k(n)}\right) = \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_{j(n)}\right)\left(x_{ik} - \bar{x}_{k(n)}\right) \qquad j,k = 1,\ldots,p$$

$$\left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right)^T \left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right)$$
$$= \left(X^T_{(n+1)} - \vec{\bar{X}}^T_{(n+1)}\right)\left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right)$$
$$= X^T_{(n+1)} X_{(n+1)} - \vec{\bar{X}}^T_{(n+1)} X_{(n+1)} - X^T_{(n+1)} \vec{\bar{X}}_{(n+1)} + \vec{\bar{X}}^T_{(n+1)} \vec{\bar{X}}_{(n+1)}$$
$$= X^T_{(n+1)} X_{(n+1)} - (n+1)\vec{\bar{X}}^T_{(n+1)} \vec{\bar{X}}_{(n+1)} - X^T_{(n+1)} \vec{\bar{X}}_{(n+1)} + \vec{\bar{X}}^T_{(n+1)} \vec{\bar{X}}_{(n+1)}$$
$$= X^T_{(n+1)} X_{(n+1)} - n\vec{\bar{X}}^T_{(n+1)} \vec{\bar{X}}_{(n+1)} - \vec{\bar{X}}^T_{(n+1)} \vec{\bar{X}}_{(n+1)} - X^T_{(n+1)} \vec{\bar{X}}_{(n+1)} + \vec{\bar{X}}^T_{(n+1)} \vec{\bar{X}}$$
$$= X^T_{(n+1)} X_{(n+1)} - n\vec{\bar{X}}^T_{(n+1)} \vec{\bar{X}}_{(n+1)} - X^T_{(n+1)} \vec{\bar{X}}_{(n+1)}$$
$$= X^T_{(n+1)} X_{(n+1)} - n\vec{\bar{X}}^T_{(n+1)} \vec{\bar{X}}_{(n+1)} - \vec{\bar{X}}^T_{(n+1)} \vec{\bar{X}}_{(n+1)}$$
$$= X^T_{(n+1)} X_{(n+1)} - (n+1)\vec{\bar{X}}^T_{(n+1)} \vec{\bar{X}}_{(n+1)}$$
$$= X^T_{(n+1)} X_{(n+1)} - (n+1)\vec{\bar{X}}_{(n+1)} \vec{\bar{X}}^T_{(n+1)}$$

En forma matricial $\left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right)^T \left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right)$ desarrollando se tiene

$$= \left\{ \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{np} \\ x_{n+1,1} & \cdots & x_{n+1,p} \end{pmatrix} - \begin{pmatrix} \bar{x}_{1(n+1)} & \cdots & \bar{x}_{p(n+1)} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \bar{x}_{1(n+1)} & \cdots & \bar{x}_{p(n+1)} \end{pmatrix} \right\}^T = \left\{ \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{np} \\ x_{n+1,1} & \cdots & x_{n+1,p} \end{pmatrix} - \begin{pmatrix} \bar{x}_{1(n+1)} & \cdots & \bar{x}_{p(n+1)} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \bar{x}_{1(n+1)} & \cdots & \bar{x}_{p(n+1)} \end{pmatrix} \right\}$$

$$= \left\{ \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{np} \\ x_{n+1,1} & \cdots & x_{n+1,p} \end{pmatrix} - \begin{pmatrix} \bar{x}_{1(n)} & \cdots & \bar{x}_{p(n)} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \bar{x}_{1(n)} & \cdots & \bar{x}_{p(n)} \end{pmatrix} \right\} = \left\{ \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{np} \\ x_{n+1,1} & \cdots & x_{n+1,p} \end{pmatrix} - \begin{pmatrix} \bar{x}_{1(n)} & \cdots & \bar{x}_{p(n)} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \bar{x}_{1(n)} & \cdots & \bar{x}_{p(n)} \end{pmatrix} \right\}^T$$

$$= \begin{bmatrix} \sum_{i=1}^{n+1}\left(x_{i1}-\bar{x}_{1(n)}\right)^2 & \sum_{i=1}^{n+1}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right) & \cdots & \cdots & \sum_{i=1}^{n+1}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{ip}-\bar{x}_{p(n)}\right) \\ \sum_{i=1}^{n+1}\left(x_{i2}-\bar{x}_{2(n)}\right)\left(x_{i1}-\bar{x}_{1(n)}\right) & \sum_{i=1}^{n+1}\left(x_{i2}-\bar{x}_{2(n)}\right)^2 & \cdots & \cdots & \sum_{i=1}^{n+1}\left(x_{i2}-\bar{x}_{2(n)}\right)\left(x_{ip}-\bar{x}_{p(n)}\right) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n+1}\left(x_{ip}-\bar{x}_{p(n)}\right)\left(x_{i1}-\bar{x}_{1(n)}\right) & \sum_{i=1}^{n+1}\left(x_{ip}-\bar{x}_{p(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right) & \cdots & \cdots & \sum_{i=1}^{n+1}\left(x_{ip}-\bar{x}_{p(n)}\right)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)^2 & \sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right) & \cdots & \cdots & \sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{ip}-\bar{x}_{p(n)}\right) \\ \sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n)}\right)\left(x_{i1}-\bar{x}_{1(n)}\right) & \sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n)}\right)^2 & \cdots & \cdots & \sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n)}\right)\left(x_{ip}-\bar{x}_{p(n)}\right) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n}\left(x_{ip}-\bar{x}_{p(n)}\right)\left(x_{i1}-\bar{x}_{1(n)}\right) & \sum_{i=1}^{n}\left(x_{ip}-\bar{x}_{p(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right) & \cdots & \cdots & \sum_{i=1}^{n}\left(x_{ip}-\bar{x}_{p(n)}\right)^2 \end{bmatrix}$$

$$= \left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)^T \left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)$$

$$= \left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right)^T \left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right)$$

$$= \left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)^T \left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)$$

$$= X_{(n)}^T X_{(n)} - n\vec{\bar{X}}_{(n)}\vec{\bar{X}}_{(n)}^T$$

Es decir se prueba que $\left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right)^T \left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right) = X_{(n)}^T X_{(n)} - n\vec{\bar{X}}_{(n)}\vec{\bar{X}}_{(n)}^T$ (13)

$$S_{(n+1)} = \frac{1}{n+1}\left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right)^T \left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right)$$

$$= \frac{1}{n+1}\left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)^T \left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)$$

$$= \frac{1}{n+1}\left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)^T \left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)\frac{n}{n}$$

$$= \frac{n}{n+1}\left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)^T\left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)\frac{1}{n}$$

$$= \frac{n}{n+1}S_{(n)}$$

Teorema 2.3 Obtención de la Matriz de covarianzas Muestral Insesgada con datos faltantes

Sea $V_{(n)} = \frac{1}{n-1}\left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right)^T\left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right)$ la matriz de varianzas y covarianzas insesgada con n observaciones conocidas. $V_{(n+1)}$ está afectada por el factor $\frac{n-1}{n}$ con respecto a la matriz de covarianzas muestral insesgada con todas las observaciones conocidas de $V_{(n)}$, es decir:

$$V_{(n+1)} = \frac{n-1}{n}V_{(n)}$$

(14)

Demostración

Se define

$$V_{(n+1)} = \frac{1}{n}\left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right)^T\left(X_{(n+1)} - \vec{\bar{X}}_{(n+1)}\right)$$
$$= \left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)^T\left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)$$
$$= \frac{n-1}{n}\frac{\left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)^T\left(X_{(n)} - \vec{\bar{X}}_{(n)}\right)}{n-1}$$
$$= \frac{n-1}{n}V_{(n)}$$

Eso equivale al factor $\frac{(n-1)}{n}$ multiplicado por la matriz de varianzas y covarianzas muestral insesgada con n observaciones conocidas

Teorema 2.4 Obtención de la Matriz de correlaciones con datos faltantes

Sea $\mathbf{R}_{(n)} = \begin{bmatrix} 1 & r_{12(n)} & \cdots & \cdots & r_{1p(n)} \\ r_{21(n)} & 1 & \cdots & \cdots & r_{2p(n)} \\ . & . & . & . & . \\ . & . & . & . & . \\ r_{p1(n)} & r_{p2(n)} & \cdots & \cdots & 1 \end{bmatrix}$ la matriz de correlaciones muestral con n observaciones, donde $r_{ij(n)}$ es la correlación entre las variable i y la variable j ,con "n" observaciones conocidas.

Si definimos

$\mathbf{R}_{(n+1)} = \begin{bmatrix} 1 & r_{12(n+1)} & \cdots & \cdots & r_{1p(n+1)} \\ r_{21(n+1)} & 1 & \cdots & \cdots & . \\ . & . & . & . & . \\ . & . & . & . & . \\ r_{p1(n+1)} & r_{p2(n+1)} & \cdots & \cdots & 1 \end{bmatrix}$ la matriz de correlaciones muestral con observaciones perdidas que previamente han sido estimadas, en cada variable por su respectiva media, entonces,

$$\mathbf{R}_{(n)} = \mathbf{R}_{(n+1)} \tag{15}$$

donde $r_{ij(n)}$ es la correlación entre las variables i y la variable j , con "$n+1$" observaciones.

Vamos a trabajar con cada correlación entre dos variables.

$$r_{12(n)} = \frac{s_{12}}{\sqrt{s_{11}s_{22}}} = \frac{\dfrac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right)}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n)}\right)^2}{n-1}}}$$

$$\frac{\dfrac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right)}{n-1}}{\dfrac{\sqrt{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)^2}}{\sqrt{n-1}}\dfrac{\sqrt{\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n)}\right)^2}}{\sqrt{n-1}}} = \frac{\dfrac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right)}{n-1}}{\dfrac{\sqrt{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n)}\right)^2}}{n-1}}$$

$$\mathrm{r}_{12(\mathrm{n})} = \frac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n)}\right)^2}}$$

(16)

Esto es la correlación muestral entre las dos primeras variables de manera que se puede generalizar al resto de pares de variables.

Asi:

$$r_{jk(n)} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} = \frac{\dfrac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)\left(x_{ik}-\bar{x}_{k(n)}\right)}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}\left(x_{ik}-\bar{x}_{k(n)}\right)^2}{n-1}}}$$

$$= \frac{\dfrac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)\left(x_{ik}-\bar{x}_{k(n)}\right)}{n-1}}{\dfrac{\sqrt{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)^2}}{\sqrt{n-1}}\dfrac{\sqrt{\sum_{i=1}^{n}\left(x_{ik}-\bar{x}_{k(n)}\right)^2}}{\sqrt{n-1}}} = \frac{\dfrac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)\left(x_{ik}-\bar{x}_{k(n)}\right)}{n-1}}{\dfrac{\sqrt{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{ik}-\bar{x}_{k(n)}\right)^2}}{n-1}} \qquad j,k=1,\ldots,p$$

$$r_{jk(n)} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} = \frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)\left(x_{ik}-\bar{x}_{k(n)}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{ik}-\bar{x}_{k(n)}\right)^2}} \qquad j,k=1,\ldots,p$$

(17)

La correlación entre las dos primeras variables para las $n+1$ observaciones es

$$r_{12(n+1)} = \frac{s_{12(n+1)}}{\sqrt{s_{11(n+1)}s_{22(n+1)}}} = \frac{\dfrac{\sum_{i=1}^{n+1}\left(x_{i1}-\bar{x}_{1(n+1)}\right)\left(x_{i2}-\bar{x}_{2(n+1)}\right)}{n}}{\sqrt{\dfrac{\sum_{i=1}^{n+1}\left(x_{i1}-\bar{x}_{1(n+1)}\right)^2}{n}}\sqrt{\dfrac{\sum_{i=1}^{n+1}\left(x_{i2}-\bar{x}_{2(n+1)}\right)^2}{n}}}$$

En esta parte se quiere probar que la diferencia de cuadrados entre las dos primeras variables para los n+1 individuos es

$$\sum_{i=1}^{n+1}\left(x_{i1}-\bar{x}_{1(n+1)}\right)\left(x_{i2}-\bar{x}_{2(n+1)}\right)=\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n+1)}\right)\left(x_{i2}-\bar{x}_{2(n+1)}\right)+\left(x_{n+1,1}-\bar{x}_{1(n+1)}\right)\left(x_{n+1,2}-\bar{x}_{2(n+1)}\right)$$

Por otro lado se sabe que el numerador de S_{11} con $n+1$individuos

$$\sum_{i=1}^{n+1}\left(x_{i1}-\bar{x}_{1(n+1)}\right)^2=\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n+1)}\right)^2+\left(x_{n+1,1}-\bar{x}_{1(n+1)}\right)^2=\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)^2 \text{ , que}$$

es equivalente al denominador del numerador de S_{11} con n observaciones más la suma de cuadrados de la $n+1$ é-sima observación perdida.

De la misma manera se desprende que:

$$\sum_{i=1}^{n+1}\left(x_{i2}-\bar{x}_{2(n+1)}\right)^2=\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n+1)}\right)^2+\left(x_{n+1,2}-\bar{x}_{2(n+1)}\right)^2=\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n)}\right)^2 \text{ , que es}$$

equivalente al numerador de S_{22}con n observaciones mas la suma de cuadrados de la n+1

é-sima observación.

Entonces

$$r_{12(n+1)}=\frac{\dfrac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n+1)}\right)\left(x_{i2}-\bar{x}_{2(n+1)}\right)+\left(x_{n+1,1}-\bar{x}_{1(n+1)}\right)\left(x_{n+1,2}-\bar{x}_{2(n+1)}\right)}{n}}{\sqrt{\dfrac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n+1)}\right)^2+\left(x_{n+1,1}-\bar{x}_{1(n+1)}\right)^2}{n}}\sqrt{\dfrac{\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n+1)}\right)^2+\left(x_{n+1,2}-\bar{x}_{2(n+1)}\right)^2}{n}}}$$

Como $x_{n+1,1}=\bar{x}_{1(n)}=\bar{x}_{1(n+1)}$ y $x_{n+1,2}=\bar{x}_{2(n)}=\bar{x}_{2(n+1)}$, entonces reemplazando estos valores se tiene:

$$r_{12(n+1)} = \frac{\dfrac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n+1)}\right)\left(x_{i2}-\bar{x}_{2(n+1)}\right)+\left(\bar{x}_{1(n)}-\bar{x}_{1(n)}\right)\left(\bar{x}_{2(n)}-\bar{x}_{2(n)}\right)}{n}}{\sqrt{\dfrac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n+1)}\right)^2+\left(\bar{x}_{1(n)}-\bar{x}_{1(n)}\right)^2}{n}}\sqrt{\dfrac{\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n+1)}\right)^2+\left(\bar{x}_{2(n)}-\bar{x}_{2(n)}\right)^2}{n}}}$$

En esta parte se está simplificando la expresión anterior

$$r_{12(n+1)} = \frac{\dfrac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n+1)}\right)\left(x_{i2}-\bar{x}_{2(n+1)}\right)}{n}}{\sqrt{\dfrac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n+1)}\right)^2}{n}}\sqrt{\dfrac{\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n+1)}\right)^2}{n}}} = \frac{\dfrac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n+1)}\right)\left(x_{i2}-\bar{x}_{2(n+1)}\right)}{n}}{\dfrac{\sqrt{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n+1)}\right)^2}}{\sqrt{n}}\dfrac{\sqrt{\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n+1)}\right)^2}}{\sqrt{n}}} =$$

$$\frac{\dfrac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n+1)}\right)\left(x_{i2}-\bar{x}_{2(n+1)}\right)}{n}}{\dfrac{\sqrt{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n+1)}\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n+1)}\right)^2}}{n}}$$

En esta parte se divide entre n individuos que determina el denominador de la primera división algebraica como n individuos que determina el denominador de la segunda división algebraica.

$$\mathrm{r}_{12(\mathrm{n}+1)} = \frac{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)\left(x_{i2}-\bar{x}_{2(n)}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_{1(n)}\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_{2(n)}\right)^2}} = \mathrm{r}_{12(\mathrm{n})}$$

Generalizando para determinar la correlación entre la j é-sima y k é-sima variable, se tiene a continuación que la matriz de correlaciones para las $n+1$ observaciones, siendo el $n+1$ é-simo individuo el individuo perdido.

A continuación se tiene que

$$r_{jk(n+1)} = \frac{s_{jk(n+1)}}{\sqrt{s_{jj(n+1)}s_{kk(n+1)}}} = \frac{\dfrac{\sum_{i=1}^{n+1}\left(x_{ij}-\bar{x}_{j(n+1)}\right)\left(x_{ik}-\bar{x}_{k(n+1)}\right)}{n+1}}{\sqrt{\dfrac{\sum_{i=1}^{n+1}\left(x_{ij}-\bar{x}_{j(n+1)}\right)^2}{n+1}}\sqrt{\dfrac{\sum_{i=1}^{n+1}\left(x_{ik}-\bar{x}_{k(n+1)}\right)^2}{n+1}}}$$

$$\sum_{i=1}^{n+1}\left(x_{ij}-\bar{x}_{j(n+1)}\right)\left(x_{ij}-\bar{x}_{j(n+1)}\right)=\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n+1)}\right)\left(x_{ik}-\bar{x}_{k(n+1)}\right)+\left(x_{n+1,j}-\bar{x}_{j(n+1)}\right)\left(x_{n+1,k}-\bar{x}_{k(n+1)}\right)$$

Por otro lado se sabe que el numerador de s_{ik} con $n+1$ observaciones, siendo la $n+1$ é-sima individuo perdido, donde

$\sum_{i=1}^{n+1}\left(x_{ij}-\bar{x}_{j(n+1)}\right)^2 = \sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n+1)}\right)^2 + \left(x_{n+1,j}-\bar{x}_{j(n+1)}\right)^2 = \sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)^2$, que es equivalente al numerador de s_{jj} con n observaciones conocidas mas la suma de cuadrados de la $n+1$ é-simo individuo perdido.

De la misma manera se desprende que:

$\sum_{i=1}^{n+1}\left(x_{ik}-\bar{x}_{k(n+1)}\right)^2 = \sum_{i=1}^{n}\left(x_{ik}-\bar{x}_{k(n+1)}\right)^2 + \left(x_{n+1,k}-\bar{x}_{k(n+1)}\right)^2$, es equivalente al numerador de s_{kk} con n observaciones mas la suma de cuadrados de la $n+1$ é-simo individuo perdido.

Ahora se sabe qué $x_{n+1,j} = \bar{x}_{j(n)} = \bar{x}_{j(n+1)}$ y $x_{n+1,k} = \bar{x}_{k(n)} = \bar{x}_{k(n+1)}$, entonces

$$r_{jk(n+1)} = \frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n+1)}\right)\left(x_{ik}-\bar{x}_{k(n+1)}\right)+\left(x_{n+1,j}-\bar{x}_{j(n+1)}\right)\left(x_{n+1,k}-\bar{x}_{k(n+1)}\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n+1)}\right)^2+\left(x_{n+1,j}-\bar{x}_{j(n+1)}\right)^2}{n}}\sqrt{\frac{\sum_{i=1}^{n}\left(x_{ik}-\bar{x}_{k(n+1)}\right)^2+\left(x_{n+1,k}-\bar{x}_{k(n+1)}\right)^2}{n}}}$$

Luego:

En esta parte se divide entre n individuos que determina tanto el denominador de la primera división algebraica como n individuos que determina tanto el denominador de la segunda división algebraica.

$$r_{jk(n+1)} = \frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n+1)}\right)\left(x_{ik}-\bar{x}_{k(n+1)}\right)+\left(\bar{x}_{j(n)}-\bar{x}_{j(n)}\right)\left(\bar{x}_{k(n)}-\bar{x}_{k(n)}\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n+1)}\right)^2+\left(\bar{x}_{j(n)}-\bar{x}_{j(n)}\right)^2}{n}}\sqrt{\frac{\sum_{i=1}^{n}\left(x_{ik}-\bar{x}_{k(n+1)}\right)^2+\left(\bar{x}_{k(n)}-\bar{x}_{k(n)}\right)^2}{n}}}$$

$$r_{jk(n+1)} = \frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n+1)}\right)\left(x_{ik}-\bar{x}_{k(n+1)}\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n+1)}\right)^2}{n}}\sqrt{\frac{\sum_{i=1}^{n}\left(x_{ik}-\bar{x}_{k(n+1)}\right)^2}{n}}}$$

Reemplazando $\bar{x}_{j(n)} = \bar{x}_{j(n+1)}$ y $\bar{x}_{k(n)} = \bar{x}_{k(n+1)}$ se tiene que

$$r_{jk(n+1)} = \frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)\left(x_{ik}-\bar{x}_{k(n)}\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)^2}{n}}\sqrt{\frac{\sum_{i=1}^{n}\left(x_{ik}-\bar{x}_{k(n)}\right)^2}{n}}}$$

$$r_{jk(n+1)} = \frac{\dfrac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)\left(x_{ik}-\bar{x}_{k(n)}\right)}{n}}{\sqrt{\dfrac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{j(n)}\right)^2}{n}}\sqrt{\dfrac{\sum_{i=1}^{n}\left(x_{ik}-\bar{x}_{k(n)}\right)^2}{n}}} = r_{jk(n)}$$

Se ha demostrado que :

$$r_{jk(n+1)} = r_{jk(n)} \qquad j,k = 1,...,p$$

Con los resultados anteriores expresados matricialmente y se tienen las matrices de correlaciones con "$n+1$" y "n" observaciones.

$$\mathbf{R}_{(n+1)} = \begin{bmatrix} 1 & r_{12(n+1)} & \cdots & \cdots & r_{1p(n+1)} \\ r_{21(n+1)} & 1 & \cdots & \cdots & . \\ . & . & . & . & . \\ . & . & . & . & . \\ r_{1p(n+1)} & r_{2p(n+1)} & \cdots & \cdots & 1 \end{bmatrix} = \begin{bmatrix} 1 & r_{12(n)} & \cdots & \cdots & r_{1p(n)} \\ r_{21(n)} & 1 & \cdots & \cdots & . \\ . & . & . & . & . \\ . & . & . & . & . \\ r_{1p(n)} & r_{2p(n)} & \cdots & \cdots & 1 \end{bmatrix} = \mathbf{R}_{(n)}$$

(18)

Se ha probado que la matriz de correlaciones muestral Insesgada es invariante al número de individuos, es decir la $n+1$ é-sima observación perdida no afecta al cálculo de la correlación entre dos variables.

CAPITULO III

EFECTO EN LA ESTIMACIÓN DE DATOS FALTANTES EN EL ANÁLISIS DE COMPONENTES PRINCIPALES

3.1 INTRODUCCIÓN

Uno de los aspectos más importantes de la estadística multivariante es el Análisis de Componentes Principales (ACP), cuyo objetivo es describir la variación total de la muestra en pocas dimensiones, es decir, se pretende reducir la dimensión del conjunto de datos original, minimizando la pérdida de información. Estas pocas dimensiones llamadas Componentes Principales son combinaciones lineales no correlacionadas de las variables originales, que en ocasiones son utilizadas como paso inicial en muchos análisis estadísticos, funcionando la técnica como análisis exploratorio de datos multivariantes.

Los Objetivos más importantes de todo análisis de Componentes Principales (ACP) son:

-Generar nuevas variables no correlaciones que puedan expresar la información contenida en el conjunto original de datos para reducir la dimensión (numero de variables) del problema que se está estudiando, como paso previo para futuros análisis.

-Eliminar, cuando sea posible, algun de las variables originales si ellas aportan poca información.

Las Nuevas variables generadas se denominan componentes Principales (CP) y poseen algunas características deseables, tales como independencia (caso de multinormalidad) y en todos los casos estas no están correlacionadas.

Cada componente sintetiza la máxima variabilidad residual contenida en los datos.

En el presente capitulo haremos una presentación del método de la obtención de las Componentes Principales como funciones de los autovectores al tener "n" observaciones conocidas y haremos una comparación al tener n+1 observaciones,

considerando sin pérdida de generalidad, la n+1 é-sima observación faltante en cada una de las variables, se desea saber qué efecto se produce imputando con la media aritmética en cada una de las variables en estudio.

3.2 ASPECTOS GENERALES

Sea $w_{1(n)},.....w_{p(n)}$: Valores propios de $\mathbf{R}_{(n)}$.

Sea $\beta_{1(n)},.....\beta_{p(n)}$: Vectores propios de $\mathbf{R}_{(n)}$

Sea $\lambda_{1(n)},.....\lambda_{p(n)}$: Valores propios de $V_{(n)}$

Sea $\theta_{1(n)},.....\theta_{p(n)}$: Vectores propios de $V_{(n)}$

Sea $\gamma_{1(n+1)},.....\gamma_{p(n+1)}$: Valores propios de $V_{(n+1)}$

Sea $\delta_{1(n+1)},.....\delta_{p(n+1)}$: Vectores propios de $V_{(n+1)}$

La matriz de correlaciones $\mathbf{R}_{(n)}$ y matriz de covarianzas V_n son matrices simétricas, es por eso que el espacio vectorial (V) de dimensión finita son los números reales.

A continuación se hace un resumen de teoremas importantes del algebra, que serán de necesidad.

3.3 RESULTADOS FUNDAMENTALES

Teorema 3. 1

Se sabe que la matriz de correlación $\mathbf{R}_{(n)}$ una matriz real y simétrica de orden pxp.

Si w es un valor propio de $\mathbf{R}_{(n)}$ en los reales, entonces w es real.

Teorema 3.2

Se sabe que la matriz de covarianzas V_n una matriz real y simétrica de orden pxp.

Si λ es un valor propio de V_n en los reales, entonces λ es real

Teorema 3.3

Se sabe que la matriz de correlación $\mathbf{R}_{(n)}$ una matriz simétrica real de orden pxp. Entonces, la matriz de correlación $\mathbf{R}_{(n)}$ tiene un vector propio real no nulo.

Teorema 3. 4

Sea la matriz de covarianzas V_n una matriz simétrica real de orden pxp. Entonces, V_n tiene un vector propio real no nulo.

Teorema 3.5 El Teorema Espectral

Se supone que (V) es un espacio de dimensión finita sobre los reales con un producto escalar definitivamente positivo. También se supone que $\dim V \geq 1$

Teorema 3. 5

Sea $R_n : R \to R$ una aplicación lineal simétrica. Entonces existe una base ortogonal de R que consta de los vectores propios de la matriz de correlación $\mathbf{R}_{(n)}$

Teorema 3.6

Sea $V_n : R \to R$ una aplicación lineal simétrica. Entonces existe una base ortogonal de R que consta de los vectores propios de la matriz de covarianzas V_n

Teorema 3.7

Sea V un espacio vectorial sobre R y sea $V_n : R \to R$ un operador.

Sean $\theta_1,....\theta_p$ vectores propios de la matriz de covarianzas V_n con valores propios $\lambda_1,.....\lambda_p$ respectivamente.

Supóngase que estos valores propios son distintos entre sí, es decir :

$\lambda_i \neq \lambda_j$ si $i \neq j$.

Entonces los autovalores $\lambda_1,.....\lambda_p$ son linealmente independientes.

Demostración:

Por inducción sobre p. Para p=1, un elemento $\lambda_1 \in V$, $\lambda_1 \neq 0$ es linealmente independiente.

Sea $p > 1$. Supóngase la relación $c_1\theta_1 + ...c_p\theta_p = 0$

Con $c_i \in R$ .Hay que probar que todo $c_i = 0$. Se multiplica la relación anterior por λ_1 para obtener

$c_1\lambda_1\theta_1 + ...c_p\lambda_1\theta_p = 0$

También se aplica V_n a la relación anterior .Por linealidad , se obtiene

$c_1\lambda_1\theta_1 + ...c_p\lambda_p\theta_p = 0$

Ahora se sustraen estas dos últimas expresiones y se obtiene

$$c_2(\lambda_2 - \lambda_1)\theta_2 + ... c_p(\lambda_p - \lambda_1)\theta_p = 0$$

Como $\lambda_j - \lambda_1 \neq 0$ para j=2.....p se concluye , por inducción , que $c_2 = ... = c_p = 0$

Volviendo a la relación original, se ve que $c_1\theta_1 = 0$, con lo que $c_1 = 0$ y el teorema queda probado.

Teorema 3.8

Sea V un espacio vectorial sobre R y sea $R_n : R \to R$ un operador. Sean $\beta_1,....\beta_p$ vectores propios de $\mathbf{R}_{(n)}$ con valores propios $w_1,.....w_p$ respectivamente.

Supóngase que estos valores propios son distintos entre sí, es decir :

$w_i \neq w_j$ si $i \neq j$.

Entonces los $w_1,.....w_p$ son linealmente independientes.

Prueba

Por inducción sobre p. Para p=1, un elemento $w_1 \in V, w_1 \neq 0$ es linealmente independiente.

Sea $p > 1$. Supóngase la relación

$$d_1\beta_1 + ... d_p\beta_p = 0$$

Con $d_i \in R$.Hay que probar que todo $d_i = 0$. Se multiplica la relación anterior por w_1 para obtener

$$d_1 w_1\beta_1 + ... d_p w_1\beta_p = 0$$

También se aplica $\mathbf{R}_{(n)}$ a la relación anterior .Por linealidad , se obtiene

$$d_1 w_1\beta_1 + ... d_p w_p\beta_p = 0$$

Ahora se sustraen estas dos últimas expresiones y se obtiene

$$d_2(w_2 - w_1)\beta_2 + ... d_p(w_p - w_1)\beta_p = 0$$

Como $w_j - w_1 \neq 0$ para j=2.....p se concluye , por inducción , que $d_2 = ... = d_p = 0$

[1]Volviendo a la relación original, se ve que $d_1 w_1 = 0$, con lo que $d_1 = 0$ y el teorema queda probado

Teorema 3.9 El polinomio característico

En este caso se realizara el polinomio característico de la matriz de covarianzas, ya que el polinomio de la matriz de correlaciones se procede de igual manera.

Sea $V_n : R \to R$ un operador lineal en el espacio vectorial en los reales de dimensión finita.

Para que un número real λ sea autovalor de V_n, es necesario y suficiente que exista $\theta \neq 0$ en R, tal que $(V_n - \lambda)\theta = 0$, es decir, que el operador $(V_n - \lambda I) : R \to R$ tenga núcleo no trivial y por tanto no sea invertible, esto sucede si, y solo si $\det(V_n - \lambda I) = 0$.

Conforme a la definición clásica de determinante, $\det(V_n - \lambda I)$ es un polinomio de grado "p" en λ, cuyo término de mayor grado es $(-1)^p \lambda^p$. Se llama polinomio característico del operador V_n y es representado por $p_{V_n}(\lambda) = \det(V_n - \lambda I)$

Las raíces reales de la ecuación algebraica $p_{V_n}(\lambda) = 0$ son las llamadas las raíces características del operador V_n. De lo dicho anteriormente, se sigue que los autovalores del operador lineal V_n son raíces características reales.

En particular son reales las raíces del polinomio característico de una matriz simétrica(o, lo que es lo mismo, de un operador autoadjunto en espacio con producto interno) pues toda matriz simétrica es semejante a una matriz diagonal que es, ciertamente, triangular.

La noción de un polinomio característico permite inferir que, si la dimensión de V es un número impar, entonces todo operador lineal $V_n : R \to R$ tiene, por lo menos un autovalor. En efecto, el polinomio característico $p_{V_n}(\lambda)$, siendo un polinomio real de grado impar, tiene por lo menos una raíz real.

No obstante la importancia de los autovalores de V_n, la determinación de los coeficientes de p_{V_n} es una tarea complicada cuando es elevado (más complicado

(3)[1] Referencia en Lages, E(1998)

aún es el cálculo de sus raíces).Sin embargo, uno de esos coeficientes es fácil de calcular: el término independiente de λ es igual a $p_{V_n}(0)$, luego es igual a $\det(V_n)$.

Por otro lado, si las raíces de p_{V_n} son $\lambda_1\lambda_2.....\lambda_p$, se tiene $p_{V_n}(\lambda) = (-1)^p(\lambda-\lambda_1).....(\lambda-\lambda_p)$.

Poniendo $w=0$ se tiene que $\det(V_n) = p_{V_n}(0) = \lambda_1\lambda_2.....\lambda_p$ .Por tanto, el determinante de V_n es igual al producto de sus raíces características.

Otro termino fácil de calcular en el polinomio $p_{V_n}(\lambda)$ es el coeficiente de λ^{p-1}. En la expresión clásica de $\det(V_n - \lambda I)$ en términos de la matriz $a = \lfloor a_{ij} \rfloor$ de V_n en una cierta base, los sumandos que contienen la potencia λ^{p-1} resultan del producto $\prod(a_{ii}-\lambda)$ de los términos de la diagonal de $a-\lambda I_n$, luego son todos de la forma $(-1)^{p-1}a_{ii}\lambda^{p-1}$.

Por lo tanto $(-1)^{p-1}\sum a_{ii}$ es el coeficiente de λ^{n-1} en el polinomio $p_{V_n}(\lambda)$.[2]

Nuevamente, la expresión $p_{V_n}(\lambda) = (-1)^p \prod_{i=1}^{p} (\lambda-\lambda_i)$ muestra que el coeficiente de λ^{p-1} es igual a $(-1)^{p-1}$ veces la suma de las raíces del polinomio $p_{V_n}(\lambda)$.Esto nos lleva a conducir que cualquier base escogida en V, la suma $\sum a_{ii}$ de los elementos de la diagonal de la matriz de V_n en esta base es la misma e igual a la suma de las raíces características de la matriz de covarianzas V_n, que es siempre un numero real.

A continuación se dan 3 casos en el que se quiere determinar los autovalores y autovectores de $R_{(n)}, V_{(n)}, V_{(n+1)}$.

CASO I

En este caso se quiere determinar los autovalores y autovectores de la matriz de correlación $R_{(n)}$

Sea $V: R^3 \to R^3$,

Sea la matriz de correlaciones con "n" observaciones conocidas

(3)[2] Referencia en Lages, E(1998)

$$R_{(n)} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix}$$

Procederemos hallar los autovalores y autovectores de la matriz de correlaciones con "n" observaciones conocidas.

Solución:

Determinando el Polinomio característico

$$p_{R_{(n)}}(w) = \det(R_{(n)} - wI) = w^3 - 3w^2 + (3 - r_{31}^2 - r_{32}^2 - r_{21}^2)w +$$
$$(r_{32}^2 + r_{31}^2 + r_{21}^2 - 2r_{12}r_{23}r_{31})$$

Hallaremos una raíz real aproximada mediante el método de Newton.[9]

Se sustituyen $p_{R_{(n)}}(w)$ y la derivada de $p_{R_{(n)}}(w)$

$$w_{i+1} = w_i - \frac{w^3 - 3w^2 + (3 - r_{31}^2 - r_{32}^2 - r_{21}^2)w + (r_{32}^2 + r_{31}^2 + r_{21}^2 - 2r_{12}r_{23}r_{31})}{3w^2 - 6w + (3 - r_{31}^2 - r_{32}^2 - r_{21}^2)}$$

dado que w se encuentra en cierto intervalo cerrado.

$$w_1 = w_0 - \frac{w_0^3 - 3w_0^2 + (3 - r_{31}^2 - r_{32}^2 - r_{21}^2)w_0 + (r_{32}^2 + r_{31}^2 + r_{21}^2 - 2r_{12}r_{23}r_{31})}{3w_0^2 - 6w_0 + (3 - r_{31}^2 - r_{32}^2 - r_{21}^2)}$$

$$w_2 = w_1 - \frac{w_1^3 - 3w_1^2 + (3 - r_{31}^2 - r_{32}^2 - r_{21}^2)w_1 + (r_{32}^2 + r_{31}^2 + r_{21}^2 - 2r_{12}r_{23}r_{31})}{3w_1^2 - 6w_1 + (3 - r_{31}^2 - r_{32}^2 - r_{21}^2)}$$

$$w_3 = w_2 - \frac{w_2^3 - 3w_2^2 + (3 - r_{31}^2 - r_{32}^2 - r_{21}^2)w_2 + (r_{32}^2 + r_{31}^2 + r_{21}^2 - 2r_{12}r_{23}r_{31})}{3w_2^2 - 6w_2 + (3 - r_{31}^2 - r_{32}^2 - r_{21}^2)}$$

Aquí se intuye que $w_1 \neq w_2 \neq w_3$, resolviendo la raíz evaluada en $p_{R_{(n)}}(w_i)$ más cercana a cero será una de las raíces reales del polinomio característico, es decir uno de los autovalores de la matriz de correlaciones. Supongamos que w_1 es dicha raíz entonces se hallaran los autovectores asociado a w_1 .

Otro método para observar entre que intervalos se encuentra los autovalores aplicaremos el *teorema del valor intermedio*.[9]

Primero analizaremos los intervalos donde la función es creciente o decreciente.

Para eso calculamos la primera derivada de $p_{R_{(n)}}(w)$ e analizamos los punto donde la función es positiva , en este caso se quieren determinar las posibles raíces del polinomio característico que vienen a ser los autovalores , es decir $w_i \geq 0$, por lo que

$$p'_{R_{(n)}}(w) = 3w^2 - 6w + (3 - r_{31}^2 - r_{32}^2 - r_{21}^2)$$

Entonces uno de los autovalores se encontraría entre n_2 y n_1, es decir $n_2 < w_2 < n_1$

$n_1 = \dfrac{6 + 2\sqrt{r_{31}^2 + r_{32}^2 + r_{21}^2}}{6}$ y $n_2 = \dfrac{6 - 2\sqrt{r_{31}^2 + r_{32}^2 + r_{21}^2}}{6}$,

$$\frac{6 - 2\sqrt{r_{31}^2 + r_{32}^2 + r_{33}^2}}{6} < w_2 < \frac{6 + 2\sqrt{r_{31}^2 + r_{32}^2 + r_{21}^2}}{6},$$

Los autovalores estarían en los intervalos siguientes $w_1 < n_2$ y $n_2 < w_2 < n_1$ y $n_1 < w_3$

es decir $w_1 < \dfrac{6 - 2\sqrt{r_{31}^2 + r_{32}^2 + r_{21}^2}}{6}$ y

$\dfrac{6 - 2\sqrt{r_{31}^2 + r_{32}^2 + r_{21}^2}}{6} < w_2 < \dfrac{6 + 2\sqrt{r_{31}^2 + r_{32}^2 + r_{21}^2}}{6}$ y $\dfrac{6 + 2\sqrt{r_{31}^2 + r_{32}^2 + r_{21}^2}}{6} < w_3$

Desarrollando los autovectores $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}$

Se tiene que $\begin{pmatrix} 1 - w_1 & r_{12} & r_{13} \\ r_{21} & 1 - w_1 & r_{23} \\ r_{31} & r_{32} & 1 - w_1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$

$$(1 - w_1)\hat{\beta}_1 + r_{12}\hat{\beta}_2 + r_{13}\hat{\beta}_3 = 0$$
$$r_{21}\hat{\beta}_1 + (1 - w_1)\hat{\beta}_2 + r_{23}\hat{\beta}_3 = 0$$
$$r_{31}\hat{\beta}_1 + r_{32}\hat{\beta}_2 + (1 - w_1)\hat{\beta}_3 = 0$$

Desarrollando $\hat{\beta}_1 = \dfrac{-r_{12}}{(1 - w_1)}\hat{\beta}_2 - \dfrac{r_{13}}{(1 - w_1)}\hat{\beta}_3$

$$(R_{32} - \frac{r_{31}r_{12}}{(1 - w_1)})\hat{\beta}_2 = \left(\frac{r_{13}^2}{(1 - w_1)} - (1 - w_1)\right)\hat{\beta}_3$$

$$\hat{\beta}_2 = \frac{\left(\dfrac{r_{13}^2}{(1 - w_1)} - (1 - w_1)\right)}{(r_{32} - \dfrac{r_{31}r_{12}}{(1 - w_1)})}\hat{\beta}_3$$

Haciendo un cambio de variable

$$\frac{\left(\frac{r_{13}^2}{(1-w_1)}-(1-w_1)\right)}{(r_{32}-\frac{r_{31}r_{12}}{(1-w_1)})}=a$$

$$\hat{\beta}_1=\left(\frac{-r_{12}}{(1-w_1)}a-\frac{r_{13}}{(1-w_1)}\right)\hat{\beta}_3$$

Luego dando una forma para el autovector asociado a w_1

$$\hat{\beta}=\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\\ \hat{\beta}_3\end{pmatrix}=\begin{bmatrix}\left(\frac{-r_{12}}{(1-w_1)}\frac{\left(\frac{r_{13}^2}{(1-w_1)}-(1-w_1)\right)}{(r_{32}-\frac{r_{31}r_{12}}{(1-w_1)})}-\frac{r_{13}}{(1-w_1)}\right)\\ \frac{\left(\frac{r_{13}^2}{(1-w_1)}-(1-w_1)\right)}{(r_{32}-\frac{r_{31}r_{12}}{(1-w_1)})}\\ 1\end{bmatrix}\hat{\beta}_3$$

Luego el vector asociado para w_1 seria

$$\begin{bmatrix}\left(\frac{-r_{12}}{(1-w_1)}\frac{\left(\frac{r_{13}^2}{(1-w_1)}-(1-w_1)\right)}{(r_{32}-\frac{r_{31}r_{12}}{(1-w_1)})}-\frac{r_{13}}{(1-w_1)}\right)\\ \frac{\left(\frac{r_{13}^2}{(1-w_1)}-(1-w_1)\right)}{(r_{32}-\frac{r_{31}r_{12}}{(1-w_1)})}\\ 1\end{bmatrix}$$

De la misma forma se procede a obtener para w_2 su autovector asociado

Desarrollando $\hat{\beta}_1=\frac{-r_{12}}{(1-w_2)}\hat{\beta}_2-\frac{r_{13}}{(1-w_2)}\hat{\beta}_3$

$$(r_{32}-\frac{r_{31}r_{12}}{(1-w_2)})\hat{\beta}_2=\left(\frac{r_{13}^2}{(1-w_2)}-(1-w_2)\right)\hat{\beta}_3$$

$$\hat{\beta}_2=\frac{\left(\frac{r_{13}^2}{(1-w_2)}-(1-w_2)\right)}{(r_{32}-\frac{r_{31}r_{12}}{(1-w_2)})}\hat{\beta}_3$$

Haciendo un cambio de variable

$$\frac{\left(\frac{r_{13}^2}{(1-w_2)}-(1-w_2)\right)}{(r_{32}-\frac{r_{31}r_{12}}{(1-w_2)})}=b$$

$$\hat{\beta}_1=\left(\frac{-r_{12}}{(1-w_2)}b-\frac{r_{13}}{(1-w_2)}\right)\hat{\beta}_3$$

Luego dando una forma para el autovector asociado a w_2

$$\hat{\beta}=\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\\ \hat{\beta}_3\end{pmatrix}=\begin{bmatrix}\left(\frac{-r_{12}}{(1-w_2)}\frac{\left(\frac{r_{13}^2}{(1-w_2)}-(1-w_2)\right)}{(r_{32}-\frac{r_{31}r_{12}}{(1-w_2)})}-\frac{r_{13}}{(1-w_2)}\right)\\ \frac{\left(\frac{r_{13}^2}{(w_2-1)}-(1-w_2)\right)}{(r_{32}-\frac{r_{31}r_{12}}{(1-w_2)})}\\ 1\end{bmatrix}\hat{\beta}_3$$

Luego el vector asociado al autovalor W_2 es

$$\begin{bmatrix} \left(\dfrac{-r_{12}}{(1-w_2)} \dfrac{\left(\dfrac{r_{13}^2}{(1-w_2)} - (1-w_2) \right)}{(r_{32} - \dfrac{r_{31}r_{12}}{(1-w_2)})} - \dfrac{r_{13}}{(1-w_2)} \right) \\ \dfrac{\left(\dfrac{r_{13}^2}{(w_2-1)} - (1-w_2) \right)}{(r_{32} - \dfrac{r_{31}r_{12}}{(1-w_2)})} \\ 1 \end{bmatrix}$$

De la misma forma se procede a obtener para w_3 su autovector asociado

Desarrollando $\hat{\beta}_1 = \dfrac{-r_{12}}{(1-w_3)}\hat{\beta}_2 - \dfrac{r_{13}}{(1-w_3)}\hat{\beta}_3$

$$(r_{32} - \frac{r_{31}r_{12}}{(1-w_2)})\hat{\beta}_2 = \left(\frac{r_{13}^2}{(1-w_2)} - (1-w_2) \right)\hat{\beta}_3$$

$$\hat{\beta}_2 = \frac{\left(\dfrac{r_{13}^2}{(1-w_3)} - (1-w_3) \right)}{(r_{32} - \dfrac{r_{31}r_{12}}{(1-w_3)})}\hat{\beta}_3$$

Haciendo un cambio de variable $\dfrac{\left(\dfrac{r_{13}^2}{(1-w_3)} - (1-w_3) \right)}{(r_{32} - \dfrac{r_{31}r_{12}}{(1-w_3)})} = c$

$$\hat{\beta}_1 = \left(\frac{-r_{12}}{(1-w_3)}c - \frac{r_{13}}{(1-w_3)} \right)\hat{\beta}_3$$

Luego dando una forma para el autovector asociado a w_3

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{bmatrix} \left(\dfrac{-r_{12}}{(1-w_3)}c - \dfrac{r_{13}}{(1-w_3)} \right) \\ \dfrac{\left(\dfrac{r_{13}^2}{(1-w_3)} - (1-w_3) \right)}{(r_{32} - \dfrac{r_{31}r_{12}}{(1-w_3)})} \\ 1 \end{bmatrix}\hat{\beta}_3$$

Donde $\hat{\beta}_3 \neq 0$

Luego el vector asociado al autovalor w_3
es:

$$\begin{bmatrix} \left(\frac{-r_{12}}{(1-w_3)}c - \frac{r_{13}}{(1-w_3)}\right) \\ \frac{\left(\frac{r_{13}^2}{(1-w_3)} - (1-w_3)\right)}{(r_{32} - \frac{r_{31}r_{12}}{(1-w_3)})} \\ 1 \end{bmatrix}$$

Luego el vector asociado al autovalor w_3 es

$$\begin{bmatrix} \left(\frac{-r_{12}}{(1-w_3)}\frac{\left(\frac{r_{13}^2}{(1-w_3)} - (1-w_3)\right)}{(r_{32} - \frac{r_{31}r_{12}}{(1-w_3)})} - \frac{r_{13}}{(1-w_3)}\right) \\ \frac{\left(\frac{r_{13}^2}{(w_3-1)} - (1-w_3)\right)}{(r_{32} - \frac{r_{31}r_{12}}{(1-w_3)})} \\ 1 \end{bmatrix}$$

R_n es diagonizable,y su forma diagonal es

$$D = \begin{bmatrix} w_1 & 0 & 0 \\ 0 & w_2 & 0 \\ 0 & 0 & w_3 \end{bmatrix}$$

Se sabe por que $\mathbf{R}_{(n+1)} = \mathbf{R}_{(n)}$

Se ha probado que la Matriz De Correlaciones Muestral Insesgada es invariante al número de individuos, es decir la $n+1$ é-sima observación perdida no afecta a las "n" individuos conocidos, los autovalores y autovectores de la matriz simétrica

$\mathbf{R}_{(n+1)}$ van a ser los mismos porque se estaría trabajando en este caso con la misma matriz.

CASO II

En este caso se quiere determinar los autovalores y autovectores de V_n

Sea $V: R^3 \to R^3$,

Sea la matriz de covarianzas con n observaciones conocidas

$$V_n = \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{bmatrix}$$

Procederemos hallar los autovalores y autovectores de la matriz de covarianzas con "n" observaciones conocidas.

Solución:

Determinando el Polinomio característico

$$p_{V_n}(\lambda) = \det(v_n - \lambda I) = \lambda^3 + (-v_{11} - v_{22} - v_{33})\lambda^2 +$$
$$(v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})\lambda +$$
$$(-v_{11}v_{22}v_{33} + v_{13}v_{31}v_{32} + v_{32}v_{23}v_{11} - v_{12}v_{23}v_{31} - v_{13}v_{21}v_{32}$$
$$+ v_{12}v_{21}v_{33})$$

Hallaremos una raíz real aproximada mediante el método de Newton.[9]

Se sustituyen $p_{V_n}(\lambda)$ y la derivada de $p_{V_n}(\lambda)$

$$\lambda_{i+1} = \lambda_i - \frac{\begin{array}{l}\lambda^3 + (-v_{11} - v_{22} - v_{33})\lambda^2 + (v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})\lambda + \\ (-v_{11}v_{22}v_{33} + v_{13}v_{31}v_{32} + v_{32}v_{23}v_{11} - v_{12}v_{23}v_{31} - v_{13}v_{21}v_{32} + v_{12}v_{21}v_{33})\end{array}}{\begin{array}{l}3\lambda^2 - 2\lambda(v_{11} + v_{22} + v_{33}) + (v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})\lambda + \\ (-v_{11}v_{22}v_{33} + v_{13}v_{31}v_{32} + v_{32}v_{23}v_{11} - v_{12}v_{23}v_{31} - v_{13}v_{21}v_{32} \\ + v_{12}v_{21}v_{33})\end{array}}$$

dado que λ se encuentra en un cierto intervalo cerrado

$$\lambda_1 = \lambda_0 - \frac{\begin{array}{l}\lambda_0^3 + (-v_{11} - v_{22} - v_{33})\lambda_0^2 + (v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})\lambda_0 + \\ (-v_{11}v_{22}v_{33} + v_{13}v_{31}v_{32} + v_{32}v_{23}v_{11} - v_{12}v_{23}v_{31} - v_{13}v_{21}v_{32} + v_{12}v_{21}v_{33})\end{array}}{3\lambda_0^2 - 2\lambda_0(v_{11} + v_{22} + v_{33}) + (v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})}$$

$$\lambda_2 = \lambda_1 - \frac{\lambda_1^3 + (-v_{11} - v_{22} - v_{33})\lambda_1^2 + (v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})\lambda_1 + (-v_{11}v_{22}v_{33} + v_{13}v_{31}v_{32} + v_{32}v_{23}v_{11} - v_{12}v_{23}v_{31} - v_{13}v_{21}v_{32} + v_{12}v_{21}v_{33})}{3\lambda_1^2 - 2\lambda_1(v_{11} + v_{22} + v_{33}) + (v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})}$$

$$\lambda_3 = \lambda_2 - \frac{\lambda_2^3 + (-v_{11} - v_{22} - v_{33})\lambda_2^2 + (v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})\lambda_2 + (-v_{11}v_{22}v_{33} + v_{13}v_{31}v_{32} + v_{32}v_{23}v_{11} - v_{12}v_{23}v_{31} - v_{13}v_{21}v_{32} + v_{12}v_{21}v_{33})}{3\lambda_2^2 - 2\lambda_1(v_{11} + v_{22} + v_{33}) + (v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})}$$

Aquí se intuye que $\lambda_1 \neq \lambda_2 \neq \lambda_3$, resolviendo la raíz evaluada en $p_{v_n}(\lambda_i)$ más cercana a cero será al menos una de las raíces reales del polinomio característico, es decir uno de los autovalores de la matriz de covarianzas. Supongamos que λ_1 es dicha raíz real entonces se hallaran los autovectores asociado a λ_1 .

Otro método para observar entre que intervalos se encuentra las raíces es decir los autovalores es aplicando el *teorema del valor intermedio*.[9]

Primero analizaremos los intervalos donde la función es creciente o decreciente.

Para eso calculamos la primera derivada de $p_{v_n}(\lambda)$ e analizamos los puntos donde la función es positiva ya que los autovalores son mayores a cero, en este caso se

quieren determinar las posibles raíces del polinomio característico que vienen a ser los autovalores , es decir $\lambda_i \geq 0$, por lo que

$$p'_{V_n}(\lambda) = 3\lambda^2 - 2\lambda(-v_{11} - v_{22} - v_{33}) + (v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})$$

Uno de los auto valores estarían entre

$$m_1 = \frac{2(v_{11} + v_{22} + v_{33}) + \sqrt{4(v_{11} + v_{22} + v_{33})^2 - 12(v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})}}{6} \text{ y}$$

$$m_2 = \frac{2(v_{11} + v_{22} + v_{33}) - \sqrt{4(v_{11} + v_{22} + v_{33})^2 - 12(v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})}}{6}$$

entonces $m_1 < \lambda_2 < m_2$

Los autovalores estarían en los intervalos siguientes $w_1 < m_2$ y $m_2 < \lambda_2 < m_1$ y $n_1 < w_3$ es

decir $w_1 < \frac{2(v_{11} + v_{22} + v_{33}) - \sqrt{4(v_{11} + v_{22} + v_{33})^2 - 12(v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})}}{6}$ y

$$\frac{2(v_{11} + v_{22} + v_{33}) - \sqrt{4(v_{11} + v_{22} + v_{33})^2 - 12(v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})}}{6} < w_2 <$$

$$\frac{2(v_{11} + v_{22} + v_{33}) + \sqrt{4(v_{11} + v_{22} + v_{33})^2 - 12(v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})}}{6}$$

y $\frac{2(v_{11}+v_{22}+v_{33})+\sqrt{4(v_{11}+v_{22}+v_{33})^2-12(v_{22}v_{33}+v_{11}v_{33}+v_{11}v_{22}-v_{12}v_{21}-v_{32}v_{23}-v_{13}v_{31})}}{6} < w_3$

Desarrollando los autovectores $\hat{\theta}=\begin{pmatrix}\hat{\theta}_1\\ \hat{\theta}_2\\ \hat{\theta}_3\end{pmatrix}$

Para λ_1

Se tiene que $\begin{pmatrix}v_{11}-\lambda_1 & v_{12} & v_{13}\\ v_{21} & v_{22}-\lambda_1 & v_{23}\\ v_{31} & v_{32} & v_{33}-\lambda_1\end{pmatrix}\begin{pmatrix}\hat{\theta}_1\\ \hat{\theta}_2\\ \hat{\theta}_3\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}$

$$(v_{11}-\lambda_1)\hat{\theta}_1+v_{12}\hat{\theta}_2+v_{13}\hat{\theta}_3=0$$
$$v_{21}\hat{\theta}_1+(v_{22}-\lambda_1)\hat{\theta}_2+v_{23}\hat{\theta}_3=0$$
$$v_{31}\hat{\theta}_1+v_{32}\hat{\theta}_2+(v_{33}-\lambda_1)\hat{\theta}_3=0$$

Desarrollando $\hat{\theta}_1=\frac{-v_{12}\hat{\theta}_2-v_{13}\hat{\theta}_3}{v_{11}-\lambda_1}$

$$(v_{32}-\frac{v_{31}v_{12}}{(V_{11}-\lambda_1)})\hat{\theta}_3=\left(-v_{22}+\frac{v_{12}^2}{(v_{11}-\lambda_1)}\right)\hat{\theta}_2$$

$$\hat{\theta}_2=\frac{(v_{32}-\frac{v_{31}v_{12}}{(v_{11}-\lambda_1)})}{\left(-v_{22}+\frac{v_{12}^2}{(v_{11}-\lambda_1)}\right)}\hat{\theta}_3$$

Haciendo un cambio de variable $\frac{(v_{32}-\frac{v_{31}v_{12}}{(v_{11}-\lambda_1)})}{\left(-v_{22}+\frac{v_{12}^2}{(v_{11}-\lambda_1)}\right)}=e$

$$\hat{\theta}_1 = \frac{(\lambda_1 - v_{33})}{v_{31}}\hat{\theta}_3 - \frac{v_{32}}{v_{31}}\frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_1)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_1)}\right)}\hat{\theta}_3$$

$$\hat{\theta}_1 = (\frac{(\lambda_1 - v_{33})}{v_{31}} - \frac{v_{32}}{v_{31}}\frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_1)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_1)}\right)})\hat{\theta}_3$$

Luego dando una forma para el autovector asociado a λ_1

$$\hat{\theta} = \begin{pmatrix}\hat{\theta}_1\\ \hat{\theta}_2\\ \hat{\theta}_3\end{pmatrix} = \begin{bmatrix} (\frac{(\lambda_1 - v_{33})}{v_{31}} - \frac{v_{32}}{v_{31}}\frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_1)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_1)}\right)}) \\ \frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_1)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_1)}\right)} \\ 1 \end{bmatrix}\hat{\theta}_3$$

Donde $\hat{\theta}_3 \neq 0$

Luego el vector asociado para λ_1 seria

$$\begin{bmatrix} (\frac{(\lambda_1 - v_{33})}{v_{31}} - \left(\frac{v_{32}}{v_{31}}\right)\frac{(v_{32} - \frac{v_{31}v_{12}}{\lambda_1 - v_{11}})}{\left(-v_{22} + \frac{v_{12}^2}{v_{11}-\lambda_1}\right)}) \\ \frac{(v_{32} - \frac{v_{31}v_{12}}{\lambda_1 - v_{11}})}{\left(-v_{22} + \frac{v_{12}^2}{v_{11}-\lambda_1}\right)} \\ 1 \end{bmatrix}$$

de la misma forma se procede a obtener para λ_2 su autovector asociado

desarrollando $\hat{\theta}_1 = \frac{-v_{12}\hat{\theta}_2 - v_{13}\hat{\theta}_3}{v_{11} - \lambda_2}$

$$(v_{32} - \frac{v_{31}v_{12}}{(v_{11} - \lambda_2)})\hat{\theta}_3 = \left(-v_{22} + \frac{v_{12}^2}{(v_{11} - \lambda_2)}\right)\hat{\theta}_2$$

$$\hat{\theta}_2 = \frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11} - \lambda_2)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11} - \lambda_2)}\right)}\hat{\theta}_3$$

Haciendo un cambio de variable $\frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11} - \lambda_2)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11} - \lambda_2)}\right)} = f$

$$\hat{\theta}_1 = \frac{(\lambda_2 - v_{33})}{V_{31}}\hat{\theta}_3 - \frac{v_{32}}{v_{31}}\frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11} - \lambda_2)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11} - \lambda_2)}\right)}\hat{\theta}_3$$

$$\hat{\theta}_1 = (\frac{(\lambda_2 - v_{33})}{v_{31}} - \frac{v_{32}}{v_{31}}\frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11} - \lambda_2)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11} - \lambda_2)}\right)})\hat{\theta}_3$$

Luego dando una forma para el autovector asociado a λ_2

$$\hat{\theta} = \begin{pmatrix}\hat{\theta}_1\\ \hat{\theta}_2\\ \hat{\theta}_3\end{pmatrix} = \begin{bmatrix} (\frac{(\lambda_2 - v_{33})}{v_{31}} - \frac{v_{32}}{v_{31}}\frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11} - \lambda_2)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11} - \lambda_2)}\right)}) \\ \frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11} - \lambda_2)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11} - \lambda_2)}\right)} \\ 1 \end{bmatrix}\hat{\theta}_3$$

Donde $\hat{\theta}_3 \neq 0$

Luego el vector asociado para λ_2 seria

$$\begin{bmatrix} (\frac{(\lambda_2 - v_{33})}{v_{31}} - \frac{v_{32}}{v_{31}} \frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_2)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_2)}\right)})) \\ \frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_2)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_2)}\right)} \\ 1 \end{bmatrix}$$

donde $\hat{\theta}_3 \neq 0$

Para λ_3

$$\begin{pmatrix} v_{11}-\lambda_3 & v_{12} & v_{13} \\ v_{21} & v_{22}-\lambda_3 & v_{23} \\ v_{31} & v_{32} & v_{33}-\lambda_3 \end{pmatrix} \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\theta}_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

de la misma forma se procede a obtener para λ_3 su autovector asociado

desarrollando $\hat{\theta}_1 = \frac{-v_{12}\hat{\theta}_2 - v_{13}\hat{\theta}_3}{(v_{11}-\lambda_3)}$

$$(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_3)})\hat{\theta}_3 = \left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_3)}\right)\hat{\theta}_2$$

$$\hat{\theta}_2 = \frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_3)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_3)}\right)}\hat{\theta}_3$$

Haciendo un cambio de variable $\dfrac{(v_{32}-\dfrac{v_{31}v_{12}}{(v_{11}-\lambda_3)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_3)}\right)}=g$

$$\hat{\theta}_1=\frac{(\lambda_3-v_{33})}{v_{31}}\hat{\theta}_3-\frac{v_{32}}{v_{31}}\frac{(v_{32}-\dfrac{v_{31}v_{12}}{(v_{11}-\lambda_3)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_3)}\right)}\hat{\theta}_3$$

$$\hat{\theta}_1=(\frac{(\lambda_3-v_{33})}{v_{31}}-\frac{v_{32}}{v_{31}}\frac{(v_{32}-\dfrac{v_{31}v_{12}}{(v_{11}-\lambda_3)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_3)}\right)})\hat{\theta}_3$$

Luego dando una forma para el autovector asociado a λ_3

$$\hat{\theta}=\begin{pmatrix}\hat{\theta}_1\\ \hat{\theta}_2\\ \hat{\theta}_3\end{pmatrix}=\begin{bmatrix}(\dfrac{(\lambda_3-v_{33})}{v_{31}}-\dfrac{v_{32}}{v_{31}}\dfrac{(v_{32}-\dfrac{v_{31}v_{12}}{(v_{11}-\lambda_3)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_3)}\right)})\\ \dfrac{(v_{32}-\dfrac{v_{31}v_{12}}{(v_{11}-\lambda_3)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_3)}\right)}\\ 1\end{bmatrix}\hat{\theta}_3$$

Donde $\hat{\theta}_3\neq 0$

Luego el vector asociado para λ_3 seria

$$\begin{bmatrix} \left(\frac{(\lambda_3 - v_{33})}{v_{31}} - \frac{v_{32}}{v_{31}} \frac{\left(v_{32} - \frac{v_{31}v_{12}}{(v_{11} - \lambda_3)}\right)}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11} - \lambda_3)}\right)}\right) \\ \frac{\left(v_{32} - \frac{v_{31}v_{12}}{(v_{11} - \lambda_3)}\right)}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11} - \lambda_3)}\right)} \\ 1 \end{bmatrix}$$

V_n es diagonizable,y su forma diagonal es

$$P = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

CASO III

Sea la matriz de covarianzas con n+1 observaciones

$$V_{n+1} = \begin{bmatrix} v_{11_{(n+1)}} & v_{12(n+1)} & v_{13(n+1)} \\ v_{21(n+1)} & v_{22_{(n+1)}} & v_{23_{(n+1)}} \\ v_{31_{(n+1)}} & v_{32_{(n+1)}} & v_{33_{(n+1)}} \end{bmatrix}$$

Procederemos hallar los autovalores y autovectores de la matriz de covarianzas con "n" observaciones conocidas.

Solución:

Determinando el Polinomio característico

$$p_{V_{n+1}}(\gamma) = \det(V_{n+1} - \gamma I) = \gamma^3 + (-v_{11_{(n+1)}} - v_{22(n+1)} - v_{33(n+1)})\gamma^2 +$$
$$(v_{22(n+1)}v_{33(n+1)} + v_{11_{(n+1)}}v_{33(n+1)} + v_{11(n+1)}v_{22_{(n+1)}} - v_{12(n+1)}v_{21_{(n+1)}} -$$
$$v_{32(n+1)}v_{23(n+1)} - v_{13(n+1)}v_{31(n+1)})\gamma +$$
$$(-v_{11(n+1)}v_{22(n+1)}v_{33(n+1)} + v_{13(n+1)}v_{31(n+1)}v_{32(n+1)} + v_{32(n+1)}v_{23(n+1)}v_{11(n+1)}$$
$$- v_{12_{(n+1)}}v_{23(n+1)}v_{31(n+1)} - v_{13(n+1)}v_{21(n+1)}v_{32(n+1)} + v_{12(n+1)}v_{21_{(n+1)}}v_{33(n+1)})$$

Un método para observar entre que intervalos se encuentra las raíces es decir los autovalores es aplicando el *teorema del valor intermedio*.[9]

Primero analizaremos los intervalos donde la función es creciente o decreciente.

Para analizar los intervalos donde la función es creciente o decreciente. Para eso calculamos la primera derivada de $p_{v_{n+1}}(\gamma)$ e analizamos los punto donde la función es positiva , en este caso se quieren determinar las posibles raíces del polinomio característico que vienen a ser los autovalores , es decir $\gamma_i \geq 0$, por lo que

$$p'_{v_{n+1}}(\gamma) = 3\gamma^2 - 2\gamma(-\left(\frac{n-1}{n}\right)v_{11} - \left(\frac{n-1}{n}\right)v_{22} - \left(\frac{n-1}{n}\right)v_{33}\ ) + \left(\frac{n-1}{n}\right)^2 (v_{22}\ v_{33}\ + v_{11}\ v_{33)} + v_{11}\ v_{22}\ -$$

$$v_{12}\ v_{21} - v_{32}\ v_{23}\ - v_{13}v_{31})$$

Luego uno de los autovalores se encontrarían entre $k_2 < \gamma_2 < k_1$

$$p'_{v_{n+1}}(\gamma) = 3\gamma^2 - 2\gamma(-v_{11_{(n+1)}} - v_{22(n+1)} - v_{33(n+1)}) + (v_{22(n+1)}v_{33(n+1)} + v_{11(n+1)}v_{33(n+1)} + v_{11(n+1)}v_{22(n+1)} -$$

$$v_{12(n+1)}v_{21(n+1)} - v_{32(n+1)}v_{23(n+1)} - v_{13(n+1)}v_{31(n+1)_(}\ )$$

$$p'_{v_{n+1}}(\gamma) = 3\gamma^2 - 2\gamma(-\left(\frac{n-1}{n}\right)v_{11} - \left(\frac{n-1}{n}\right)v_{22} - \left(\frac{n-1}{n}\right)v_{33}\ ) + \left(\frac{n-1}{n}\right)^2 (v_{22}\ v_{33}\ + v_{11}\ v_{33)} + v_{11}\ v_{22}\ -$$

$$v_{12}\ v_{21} - v_{32}\ v_{23}\ - v_{13}v_{31})$$

$$k_1 = \frac{2\left(\frac{n-1}{n}\right)(v_{11} + v_{22} + v_{33}) + \left(\frac{n-1}{n}\right)\sqrt{4(v_{11} + v_{22} + v_{33})^2 - 12(v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})}}{6} \geq 0$$

$$k_1 = \left(\frac{n-1}{n}\right)\frac{2(v_{11} + v_{22} + v_{33}) + \sqrt{4(v_{11} + v_{22} + v_{33})^2 - 12(v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})}}{6}$$ y

$$k_1 = \left(\frac{n-1}{n}\right)m_1$$

$$k_2 = \frac{2\left(\frac{n-1}{n}\right)(v_{11} + v_{22} + v_{33}) - \left(\frac{n-1}{n}\right)\sqrt{4(v_{11} + v_{22} + v_{33})^2 - 12(v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})}}{6}$$

$$k_2 = \left(\frac{n-1}{n}\right)\frac{2(v_{11} + v_{22} + v_{33}) - \sqrt{4(v_{11} + v_{22} + v_{33})^2 - 12(v_{22}v_{33} + v_{11}v_{33} + v_{11}v_{22} - v_{12}v_{21} - v_{32}v_{23} - v_{13}v_{31})}}{6}$$

$$k_2 = \left(\frac{n-1}{n}\right)m_2$$

Los autovalores estarían en los intervalos siguientes $\gamma_1 < \left(\frac{n-1}{n}\right)m_2$ y $\left(\frac{n-1}{n}\right)m_2 < \gamma_2 < \left(\frac{n-1}{n}\right)m_1$ y $\left(\frac{n-1}{n}\right)m_1 < \gamma_3$, entonces cada uno de los nuevos autovalores

estarían afectados por el factor $\left(\frac{n-1}{n}\right)$, es decir $\left(\frac{n-1}{n}\right)\lambda_1 < \left(\frac{n-1}{n}\right)m_2$ y $\left(\frac{n-1}{n}\right)m_2 < \left(\frac{n-1}{n}\right)\lambda_2 < \left(\frac{n-1}{n}\right)m_1$ y $\left(\frac{n-1}{n}\right)m_1 < \left(\frac{n-1}{n}\right)\lambda_3$,

Desarrollando los autovectores

$$\hat{\delta} = \begin{pmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \\ \hat{\delta}_3 \end{pmatrix}$$

Para $\gamma_1 = \left(\frac{n-1}{n}\right)\lambda_1$

$$\left(\frac{n-1}{n}\right)\begin{pmatrix} v_{11}-\lambda_1 & v_{12} & v_{13} \\ v_{21} & v_{22}-\lambda_1 & v_{23} \\ v_{31} & v_{32} & v_{33}-\lambda_1 \end{pmatrix}\begin{pmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \\ \hat{\delta}_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\left(\frac{n-1}{n}\right)$$

Se tiene que

$$\frac{n-1}{n}\left((v_{11}-\lambda_1)\hat{\delta}_1 + v_{12}\hat{\delta}_2 + v_{13}\hat{\delta}_3\right) = 0\left(\frac{n-1}{n}\right)$$

$$\left(v_{21}\hat{\delta}_1 + (v_{22}-\lambda_1)\hat{\delta}_2 + v_{23}\hat{\delta}_3\right)\left(\frac{n-1}{n}\right) = 0\left(\frac{n-1}{n}\right)$$

$$\left(v_{31}\hat{\delta}_1 + v_{32}\hat{\delta}_2 + (v_{33}-\lambda_1)\hat{\delta}_3\right)\left(\frac{n-1}{n}\right) = 0\left(\frac{n-1}{n}\right)$$

Desarrollando $\hat{\delta}_1 = \frac{-v_{12}\hat{\delta}_2 - v_{13}\hat{\delta}_3}{v_{11}-\lambda_1}$

$$(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_1)})\hat{\delta}_3 = \left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_1)}\right)\hat{\delta}_2$$

$$\hat{\delta}_2 = \frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_1)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_1)}\right)}\delta_3$$

Haciendo un cambio de variable $\dfrac{(v_{32}-\dfrac{v_{31}v_{12}}{(v_{11}-\lambda_1)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_1)}\right)}=h$

$$\hat{\delta}_1=\frac{(\lambda_1-v_{33})}{v_{31}}\hat{\delta}_3-\frac{v_{32}}{v_{31}}\frac{(v_{32}-\dfrac{v_{31}v_{12}}{(v_{11}-\lambda_1)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_1)}\right)}\hat{\delta}_3$$

$$\hat{\delta}_1=(\frac{(\lambda_1-v_{33})}{v_{31}}-\frac{v_{32}}{v_{31}}\frac{(v_{32}-\dfrac{v_{31}v_{12}}{(v_{11}-\lambda_1)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_1)}\right)})\hat{\delta}_3$$

Luego dando una forma para el autovector asociado a γ_1

$$\hat{\delta}=\begin{pmatrix}\hat{\delta}_1\\ \hat{\delta}_2\\ \hat{\delta}_3\end{pmatrix}=\begin{bmatrix}(\dfrac{(\lambda_1-v_{33})}{v_{31}}-\dfrac{v_{32}}{v_{31}}\dfrac{(v_{32}-\dfrac{v_{31}v_{12}}{(v_{11}-\lambda_1)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_1)}\right)})\\ \dfrac{(v_{32}-\dfrac{v_{31}v_{12}}{(v_{11}-\lambda_1)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_1)}\right)}\\ 1\end{bmatrix}\hat{\delta}_3$$

Donde $\hat{\delta}_3\neq 0$

Luego el vector asociado para γ_1 seria

$$\begin{bmatrix} (\frac{(\lambda_1 - v_{33})}{v_{31}} - \frac{v_{32}}{v_{31}} \frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11} - \lambda_1)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11} - \lambda_1)}\right)})) \\ \frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11} - \lambda_1)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11} - \lambda_1)}\right)} \\ 1 \end{bmatrix}$$

de la misma forma se procede a obtener para γ_2 su autovector asociado

desarrollando $\hat{\delta}_1 = \frac{-v_{12}\hat{\delta}_2 - v_{13}\hat{\delta}_3}{V_{11} - \lambda_2}$

$$(v_{32} - \frac{v_{31}V_{12}}{(v_{11} - \lambda_2)})\hat{\delta}_3 = \left(-v_{22} + \frac{v_{12}^2}{(v_{11} - \lambda_2)}\right)\hat{\delta}_2$$

$$\hat{\delta}_2 = \frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11} - \lambda_2)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11} - \lambda_2)}\right)}\hat{\delta}_3$$

Haciendo un cambio de variable $\frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11} - \lambda_2)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11} - \lambda_2)}\right)} = i$

$$\hat{\delta}_1 = \frac{(\lambda_2 - v_{33})}{v_{31}}\hat{\delta}_3 - \frac{v_{32}}{v_{31}}\frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_2)})}{\left(-V_{22} + \frac{V_{12}^2}{(V_{11}-\lambda_2)}\right)}\hat{\delta}_3$$

$$\hat{\delta}_1 = (\frac{(\lambda_2 - v_{33})}{(v_{31})} - \frac{(v_{32})}{v_{(31)}}\frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_2)})}{\left(-v_{22} + \frac{v_{(12)}^2}{(v_{11}-\lambda_2)}\right)})\hat{\delta}_3$$

Luego dando una forma para el autovector asociado a γ_3

$$\hat{\delta} = \begin{pmatrix}\hat{\delta}_1\\ \hat{\delta}_2\\ \hat{\delta}_3\end{pmatrix} = \begin{bmatrix} (\frac{(\lambda_2 - v_{33})}{v_{31}} - \frac{v_{32}}{v_{31}}\frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_2)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_2)}\right)})) \\ \frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_2)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_2)}\right)} \\ 1 \end{bmatrix}\hat{\delta}_3$$

Donde $\hat{\theta}_3 \neq 0$

Luego el vector asociado para γ_2 seria

$$\begin{bmatrix} (\frac{(\lambda_2 - v_{33})}{v_{31}} - \frac{v_{32}}{V_{31}}\frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_2)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_2)}\right)})) \\ \frac{(v_{32} - \frac{v_{31}v_{12}}{(v_{31}-\lambda_2)})}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_2)}\right)} \\ 1 \end{bmatrix}$$

Siendo $\hat{\delta}_3 \neq 0$

El autovector asociado al autovalor

$$\left(\frac{n-1}{n}\right)\lambda_2 = \gamma_2$$

$$\text{seria} \begin{bmatrix} \left(\frac{(\lambda_2 - v_{33})}{v_{31}} - \frac{v_{32}}{v_{31}} \frac{\left(v_{32} - \frac{v_{31}v_{12}}{(V_{11}-\lambda_2)}\right)}{\left(-V_{22} + \frac{V_{12}^2}{(V_{11}-\lambda_2)}\right)}\right) \\ \frac{\left(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_2)}\right)}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_2)}\right)} \\ 1 \end{bmatrix}$$

De la misma forma se procede a obtener el autovector asociado a su autovalor

$$\gamma_3 = \left(\frac{n-1}{n}\right)\lambda_3$$

Se tiene que $\left(\frac{n-1}{n}\right)\begin{pmatrix} v_{11}-\lambda_3 & v_{12} & v_{13} \\ v_{21} & v_{22}-\lambda_3 & v_{23} \\ v_{31} & v_{32} & v_{33}-\lambda_3 \end{pmatrix}\begin{pmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \\ \hat{\delta}_3 \end{pmatrix} = \left(\frac{n-1}{n}\right)\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$

$$\frac{n-1}{n}\left((v_{11}-\lambda_3)\hat{\delta}_1 + v_{12}\hat{\delta}_2 + v_{13}\hat{\delta}_3\right) = 0\left(\frac{n-1}{n}\right)$$

$$\left(v_{21}\hat{\delta}_1 + (v_{22}-\lambda_3)\hat{\delta}_2 + v_{23}\hat{\delta}_3\right)\left(\frac{n-1}{n}\right) = 0\left(\frac{n-1}{n}\right)$$

$$\left(V_{31}\hat{\delta}_1 + V_{32}\hat{\delta}_2 + (V_{33}-\lambda_3)\hat{\delta}_3\right)\left(\frac{n-1}{n}\right) = 0\left(\frac{n-1}{n}\right)$$

Entonces

$$\begin{pmatrix} v_{11}-\lambda_3 & v_{12} & v_{13} \\ v_{21} & v_{22}-\lambda_3 & v_{23} \\ v_{31} & v_{32} & v_{33}-\lambda_3 \end{pmatrix}\begin{pmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \\ \hat{\delta}_3 \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

de la misma forma se procede a obtener para γ_3 su autovector asociado

desarrollando $\hat{\delta}_1=\dfrac{-v_{12}\hat{\delta}_2-v_{13}\hat{\delta}_3}{(v_{11}-\lambda_3)}$

$$(v_{32}-\frac{v_{31}v_{12}}{(V_{11}-\lambda_3)})\hat{\delta}_3=\left(-v_{22}+\frac{V_{12}^2}{(V_{11}-\lambda_3)}\right)\hat{\delta}_2$$

$$\hat{\delta}_2=\frac{(V_{32}-\dfrac{V_{31}V_{12}}{(V_{11}-\lambda_3)})}{\left(-V_{22}+\dfrac{V_{12}^2}{(V_{11}-\lambda_3)}\right)}\hat{\delta}_3$$

Haciendo un cambio de variable $\dfrac{(v_{32}-\dfrac{v_{31}v_{12}}{(v_{11}-\lambda_3)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_3)}\right)}=j$

$$\hat{\delta}_1=\frac{(\lambda_3-v_{33})}{v_{31}}\hat{\delta}_3-\frac{v_{32}}{v_{31}}\frac{(v_{32}-\dfrac{v_{31}v_{12}}{(V_{11}-\lambda_3)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_3)}\right)}\hat{\delta}_3$$

$$\hat{\delta}_1=(\frac{(\lambda_3-v_{33})}{v_{31}}-\frac{v_{32}}{v_{31}}\frac{(v_{(32)}-\dfrac{v_{31}v_{12}}{(v_{11}-\lambda_3)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_3)}\right)})\hat{\delta}_3$$

Luego dando una forma para el autovector asociado a γ_3

$$\hat{\delta}=\begin{pmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \\ \hat{\delta}_3 \end{pmatrix}=\begin{bmatrix} (\dfrac{(\lambda_3-v_{33})}{v_{31}}-\dfrac{v_{32}}{v_{31}}\dfrac{(v_{32}-\dfrac{v_{31}v_{12}}{(v_{11}-\lambda_3)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_3)}\right)}) \\ \dfrac{(v_{32}-\dfrac{v_{31}v_{12}}{(v_{11}-\lambda_3)})}{\left(-v_{22}+\dfrac{v_{12}^2}{(v_{11}-\lambda_3)}\right)} \\ 1 \end{bmatrix}\hat{\delta}_3$$

Donde $\hat{\delta}_3 \neq 0$

Luego el autovector asociado al autovalor

$\left(\frac{n-1}{n}\right)\lambda_3 = \gamma_3$

seria

$$\begin{bmatrix} \left(\frac{(\lambda_3 - v_{33})}{v_{31}} - \frac{v_{32}}{v_{31}} \frac{\left(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_3)}\right)}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_3)}\right)}\right) \\ \frac{\left(v_{32} - \frac{v_{31}v_{12}}{(v_{11}-\lambda_3)}\right)}{\left(-v_{22} + \frac{v_{12}^2}{(v_{11}-\lambda_3)}\right)} \\ 1 \end{bmatrix}$$

Es decir $\hat{\theta}_1 = \hat{\delta}_1, \hat{\theta}_2 = \hat{\delta}_2, \hat{\theta}_3 = \hat{\delta}_3$, es decir los autovectores se mantienen

Se ha probado que los autovectores son invariantes solo este caso al número de observaciones en la Matriz de Covarianzas.

La Matriz diagonalizable de V_{n+1} es:

$$Q = \begin{bmatrix} \gamma_1 & 0 & 0 \\ 0 & \gamma_2 & 0 \\ 0 & 0 & \gamma_3 \end{bmatrix} = \begin{bmatrix} \left(\frac{n-1}{n}\right)\lambda_1 & 0 & 0 \\ 0 & \left(\frac{n-1}{n}\right)\lambda_2 & 0 \\ 0 & 0 & \left(\frac{n-1}{n}\right)\lambda_3 \end{bmatrix}$$

Generalizando setienen los siguientes teoremas.

Teorema 3. 10

GENERALIZACIÓN DE LOS AUTOVECTORES Y AUTOVALORES DE LAS MATRIZ $R_{(n)}$

Sea V un espacio Vectorial sobre el campo de los reales. Los autovalores y autovectores se llegan a obtener mediante la ecuación $\left|R_{(n)} - wI\right| = 0$.

Supóngase que existe una base $\{\beta_{1(n)}, \beta_{2(n)}, \beta_{3(n)}.....\beta_{p(n)}\}$ de V que consta de vectores propios de la matriz $R_{(n)}$ con valores propios $W_{1(n)}, ...W_{p(n)}$ respectivamente. Entonces, la matriz asociada con R_n respecto a esta base es la matriz diagonal.

$$D = \begin{bmatrix} w_{1(n)} & 0 & .. & .. & 0 \\ 0 & w_{2(n)} & ... & ... & 0 \\ ... & & w_{j(n)} & & ... \\ .... & & & ... & ... \\ 0 & 0 & ... & ... & w_{p(n)} \end{bmatrix}$$

[3]Al determinar $n+1$observaciones los autovalores y autovectores se llegan a obtener mediante la ecuación $|R_{(n+1)} - wI| = 0$, pero por (18) se ha comprobado que $R_{(n+1)} = R_{(n)}$.

Resolver $|R_{(n)} - wI| = 0$es equivalente a resolver $|R_{(n+1)} - wI| = 0$

por lo que $\{\beta_{1(n+1)}, \beta_{2(n+1)}, \beta_{3(n+1)}.....\beta_{p(n+1)}\}$ y $\{W_{1(n+1)}, ...W_{p(n+1)}\}$son los autovectores y autovalores de la matriz $R_{(n+1)}$, cumpliéndose para un " j " cualquiera $(1 \le j \le p)$ que $\{\beta_{j(n)} = \beta_{j(n+1)},\}$ y $\{W_{j(n)} = W_{j(n+1)}\}$

Es decir los autovalores en la matriz de correlaciones no es afectada al tener $n+1$observaciones, los autovectores es invariante al número de observaciones.

(19)

Teorema 3. 11

GENERALIZACIÓN DE LOS AUTOVECTORES Y AUTOVALORES DE LAS MATRIZ V_n Y V_{n+1}

(13)[3] Referencia en Lan S.Teorema 11(1976)

Sea V un espacio Vectorial sobre el campo de los reales. Los autovalores y autovectores se llegan a obtener mediante la ecuación $\left|V_{(n)} - \lambda I\right| = 0$. Supóngase que existe una base $\left\{\theta_{1(n)}, \theta_{2(n)}, \theta_{3(n)} \ldots\ldots \theta_{p(n)}\right\}$ de V que consta de vectores propios de la matriz V_n con valores propios $\lambda_{1(n)}, \ldots\ldots \lambda_{p(n)}$ respectivamente.

Entonces, la matriz asociada con $V_{(n)}$ respecto a esta base es la matriz diagonal.

$$P = \begin{bmatrix} \lambda_{1(n)} & 0 & .. & .. & 0 \\ 0 & \lambda_{2(n)} & ... & ... & 0 \\ ... & & \lambda_{j(n)} & & ... \\ .... & & & ... & ... \\ 0 & 0 & ... & ... & \lambda_{p(n)} \end{bmatrix}$$

Al determinar para n+1 observaciones los autovalores y autovectores se determina mediante la ecuación $\left|V_{(n+1)} - \gamma I\right| = 0$, pero por (14) se ha probado que $V_{(n+1)} = \left(\frac{n-1}{n}\right) V_{(n)}$.

Resolver $\left|V_{(n)} - \lambda I\right| = 0$ es equivalente a resolver $\left|V_{(n+1)} - \gamma I\right| = 0$ por lo que $\left\{\theta_{1(n+1)}, \theta_{2(n+1)}, \theta_{3(n+1)} \ldots\ldots \theta_{p(n+1)}\right\}$ y $\left\{\gamma_{1(n+1)} = \left(\frac{n-1}{n}\right)\lambda_{1(n)}, \ldots\ldots, \gamma_{p(n+1)} = \left(\frac{n-1}{n}\right)\lambda_{p(n)}\right\}$ son los autovectores y autovalores de la matriz $V_{(n+1)}$, cumpliéndose para un "j" cualquiera ($1 \le j \le p$) que $\left\{\theta_{j(n+1)} = \theta_{j(n)},\right\}$ y $\left\{\gamma_{j(n+1)} = \left(\frac{n-1}{n}\right)\lambda_{j(n)}\right\}$

Es decir los autovalores en la matriz de covarianzas de $V_{(n+1)}$ es afectada en $\left(\frac{n-1}{n}\right)$

en las n observaciones conocidas respectivamente, los autovectores es invariante al número de observaciones.

Entonces, la matriz asociada con V_{n+1} respecto a esta base es la matriz diagonal.

$$Q=\begin{bmatrix} \frac{(n-1)}{n}\lambda_{1(n)} & 0 & .. & .. & 0 \\ 0 & \frac{(n-1)}{n}\lambda_{2(n)} & ... & ... & 0 \\ ... & & \frac{(n-1)}{n}\lambda_{j(n)} & & ... \\ .... & & & ... & ... \\ 0 & 0 & ... & ... & \frac{(n-1)}{n}\lambda_{p(n)} \end{bmatrix}$$

El teorema del valor intermedio me indica en que intervalos se encuentra los autovalores de $R_{(n)}, V_{(n)}, V_{(n+1)}$.

Luego para p=3, se hallaron los autovalores y autovectores de $R_{(n)}, V_{(n)}, V_{(n+1)}$, demostrando que los autovectores de $V_{(n)}$ y $V_{(n+1)}$ son invariantes al numero de observaciones. Se sabe que la matriz diagonal esta en función de sus autovalores de

$R_{(n)}, V_{(n)}, V_{(n+1)}$, se demuestra que la matriz asociada con $V_{(n+1)}$ es la matriz diagonal

$$Q=\left(\frac{n-1}{n}\right)P$$

(20)

CAPÍTULO IV

APLICACIONES

4.1 Aspectos generales

En este capítulo se pretende comparar los resultados de hacer el análisis de componentes principales con observaciones faltantes, las que han sido estimadas mediante la media aritmética de las observaciones conocidas de la respectiva variable, para lo cual se usó el software R[12] con la siguiente metodología.

1. Se simula una muestra aleatoria de tamaño 30 para 3, 4, 5 y 6 variables, donde todas las observaciones son conocidas. Se realiza el análisis de componentes principales en cada caso, usando la matriz de correlaciones, prestando particular importancia a la obtención de sus autovalores y los autovectores.

2. Se simula una muestra aleatoria de tamaño 31 para 3, 4, 5 y 6, con datos faltantes en las diversas variables, los que en cada caso se estiman por su respectiva media. Se obtiene la matriz de correlaciones de la muestra y se realiza el análisis de componentes principales sobre dichas matrices, prestando particular importancia a la obtención de sus autovalores y autovectores.

3. Se comparan los resultados de (1) y (2)

4. Se repiten los pasos anteriores usando las matrices de covarianzas sesgada e insesgada respectivamente.

   Cabe indicar que se realizaron otras simulaciones que llevaron a los mismos resultados.

4.2 Análisis de Componentes Principales con 3 variables

Caso 1A: Análisis comparativo con 3 variables usando la matriz de correlaciones

Parte 1: <u>Con 30 observaciones todas conocidas</u>

Paso 1 La matriz de datos $X_{30\,x\,3}$

(Ver anexo)

Paso 2. La matriz de correlaciones para 3 variables para la muestra de tamaño 30

$$R_n = \begin{bmatrix} 1 & -0.03159465 & 0.00274303\ 6 \\ -0.03159465 & 1 & -0.01814251\ 9 \\ 0.00274303\ 6 & -0.01814251\ 9 & 1 \end{bmatrix}$$

Paso 3 Identificación de los elementos del análisis de componentes principales, es definir la ecuación de la primera componente principal

$$Y_1 = 0.6107307\left(\frac{X_1 - 0.5426}{\sqrt{0.01531164}}\right) - 0.6950736\left(\frac{X_2 - 0.5195}{\sqrt{0.0201138}}\right) + 0.3793161\left(\frac{X_3 - 0.5487}{\sqrt{0.01006425}}\right)$$

$$Y_1 = 4.9356\,X_1 - 4.90\,X_2 + 3.7810\,X_3 - 7.2988$$

Parte 2: Con 31 observaciones donde la 31 observación es perdida

Paso 4. Considerar los mismos datos donde la 31 ava observación se estima con la media de las 30 observaciones conocidas.

La matriz de datos X_{31x3}

(Ver anexo)

Paso 5 Obtener la matriz de correlaciones de la matriz con 31 datos

$$R_{n+1} = \begin{bmatrix} 1 & -0.03159465 & 0.00274303\ 6 \\ -0.03159465 & 1 & -0.01814251\ 9 \\ 0.00274303\ 6 & -0.01814251\ 9 & 1 \end{bmatrix}$$

Paso 6 Identificación de los elementos del análisis de componentes principales, es decir definir la ecuación de la primera componente principal

$$Y_1 = 0.6107307\left(\frac{X_1 - 0.5426}{\sqrt{\left(\frac{29}{30}\right)0.01531164}}\right) - 0.6950736\left(\frac{X_2 - 0.5195}{\sqrt{\left(\frac{29}{30}\right)0.0201138}}\right)$$

$$+ 0.3793161\left(\frac{X_3 - 0.5487}{\sqrt{\left(\frac{29}{30}\right)0.01006425}}\right)$$

$$Y_1 = 5.01996\ X_1 - 4.984772\ X_2 + 3.8457\ X_3 - 9.1591219$$

Paso 7

Obtener los autovalores y autovectores de la matriz de correlaciones del paso 2

En el cuadro comparativo Nº1 se observa que los autovalores y autovectores teniendo 30 datos conocidos es lo mismo con 31 observaciones donde la 31ava observación perdida es estimada por la media de las 30 observaciones conocidas.

Resumen :Cuadro Comparativo Nº1

Vector de Medias		*Autovalores*	*Matriz de Autovectores*		
	0,5426	1,03765	0,6107	-0,4954	-0,61768
n=30	0,5195	0,99763	-0,5951	0,03824	0,71792
	0,5487	0,9647	0,3793	0,86779	0,321015
	0,5426	1,03765	0,6107	-0,4954	-0,61768
n=31	0,5195	0,99763	-0,6951	0,0382	0,71792
	0,5487	0,9647	0,3793	0,8678	0,3210156

-Comparando estos resultados se obtiene que los elementos del análisis de componentes principales usando la matriz de correlaciones $\mathbf{R}_{(n)}$ y al ser comparado con los elementos del análisis componentes principales usando la matriz de

correlaciones $\mathbf{R}_{(n+1)}$, son casi exactamente iguales las componentes con las 30 observaciones conocidas.

CASO 1B: Análisis comparativo con 3 variables usando la matriz de covarianzas sesgadas

Parte 1: <u>Con 30 observaciones todas conocidas</u>

Paso 1 La matriz de datos X_{30x3} (Ver anexo)

Paso 2. La matriz de covarianzas sesgada para 3 variables para la muestra de tamaño 30

$$S_n = \begin{bmatrix} 0.01480125 & -0.00053597 & 0.00003291 6 \\ -0.00053597 & 0.01944336 & -0.00024952 \\ 0.00003291 6 & -0.0002495 & 0.0097287 \end{bmatrix}$$

Parte 2: Con 31 observaciones donde la 31 observación es perdida

Paso 3. Considerar la mismos datos donde la 31 ava observación se estima con la media de las 30 observaciones conocidas, resultando la misma matriz de datos estimada las observaciones perdidas con la media.

Estos datos han sido simulados con distribución normal estándar.

Paso 4 Obtener la matriz de covarianzas de la matriz con 31 observaciones.

$$S_{n+1} = \begin{bmatrix} 0.014307 & -0.000518 & 0.0000318 \\ -0.000518 & 0.01879 & -0.000241 \\ 0.0000318 & -0.000241 & 0.0094044 \end{bmatrix}$$

Paso 5

En resumen:

Comparar los resultados del paso 2 con los del paso 4, se demuestra que la

la varianza muestral sesgada al tener la 31ava observación perdida queda afectado por el factor $(\frac{n}{n+1})$ de la varianza muestral sesgada al tener 30 observaciones conocidas.

Resumen :Cuadro Comparativo Nº2

Vector de Medias		*Autovalores*	*Matriz de Autovectores*		
	0,5426	0,01951	0,1132	0,99353	-0,0038
n=30	0,5195	0,01474	-0,9932	0,11327	0,0254
	0,5487	0,00972	0,0257	0,00089	0,9996
	0,5426	0,018861	0,1132	0,99353	-0,0038
n=31	0,5195	0,014248	-0,9932	0,11327	0,0254
	0,5487	0,009398	0,0257	0,00089	0,9996

Del cuadro comparativo Nº2, se demuestra que los autovalores con la 31 ava observación perdida que es estimada por la media de la 30 observaciones conocidas, disminuyen en $(\frac{n-1}{n})$ de los autovalores al tener los 30 datos conocidos. Los autovectores es invariante al número de observaciones.

CASO 1C: Análisis comparativo con 3 variables usando la matriz de covarianzas insesgadas

Parte 1: <u>Con 30 observaciones todas conocidas</u>

Paso 1 La matriz de datos X_{30x3}

(Ver anexo)

Estos datos han sido simulados con distribución normal estándar.

Paso 2. . La matriz de covarianzas insesgada para 3 variables para la muestra de tamaño 30

$$V_n = \begin{bmatrix} 0.01531164 & -0.00055446 & 0.00003405 \\ -0.00005545 & 0.020113825 & -0.0001693 \\ 0.00003405 & -0.000258128 & 0.01006425 \end{bmatrix},$$

Paso 3 Identificación de los elementos del análisis de componentes principales, es decir definir la ecuación de la primera componente principal

$$Y_1 = 0.1132\,X_1 - 0.9932\,X_2 + 0.00257\,X_3$$

Parte 2: Con 31 observaciones donde la 31 observación es perdida

Paso 4. Considerar los mismos datos donde la 31 ava observación se estima con la media de las 30 observaciones conocidas, resultando la misma matriz de datos estimada las observaciones perdidas con la media

La matriz de datos X_{31x3}

Estos datos han sido simulados con distribución normal estándar.

(Ver anexo)

Paso 5 Obtener la matriz de covarianzas de la matriz con 31 datos

$$V_{n+1} = \begin{bmatrix} 0.014800125 & -0.000535979 & 0.0000329162 \\ -0.000535979 & 0.0194433644 & -0.000249524 \\ 0.0000329162 & -0.000249524 & 0.009728775 \end{bmatrix}$$

Paso 6 Identificación de los elementos del análisis de componentes principales, es decir definir la ecuación de la primera componente principal

$$Y_1 = 0.1132X_1 - 0.9932X_2 + 0.00257X_3$$

Paso 7

Resumen : Cuadro Comparativo Nº3

	Vector de Medias	*Autovalores*	*Matriz de Autovectores*
n=30	0,5426	0,02018	$\theta = \begin{bmatrix} 0.1132 & 0.99356 & -0.003792 \\ -0.9932 & 0.1132 & 0.0254507 \\ 0.00257 & 0.00088 & 0.9996688 \end{bmatrix}$
	0,5195	0,01525	
	0,5487	0,01006	
n=31	0,5426	0,019511	$\delta = \begin{bmatrix} 0.1132 & 0.99356 & -0.003792 \\ -0.9932 & 0.1132 & 0.0254507 \\ 0.00257 & 0.00088 & 0.9996688 \end{bmatrix}$
	0,5195	0,014740	
	0,5487	0,009723	

En el cuadro comparativo Nº3 los autovalores teniendo 31 observaciones donde la 31ava observación perdida es estimada mediante la media de las 30 observaciones

conocidas, disminuye aproximadamente en $(n\text{-}1/n)$ de los autovalores de las 30 observaciones conocidas.

- Se demuestra que los autovectores teniendo los 30 datos conocidos

es igual a los autovectores con 31 observaciones donde la 31 observación es estimado mediante la media de las 30 observaciones conocidas, $\theta_{1(n)} = \delta_{1(n+1)}, \theta_{2(n)} = \delta_{2(n+1)}, \theta_{3(n)} = \delta_{3(n+1)}$.

-Se demuestra que

$$\gamma_{1(n+1)=}\left(\frac{29}{30}\right)\lambda_{1(n)}, \gamma_{2(n+1)=}\left(\frac{29}{30}\right)\lambda_{2(n)}, \gamma_{3(n+1)=}\left(\frac{29}{30}\right)\lambda_{3(n)}$$

En cambio los autovalores, al comparar el paso 3 con el paso 7

Con 31 observaciones, al tener la 31ava observación perdida, que es estimada por la media de las 30 conocidas, disminuye en $(\frac{n-1}{n})$ al ser comparada con los

autovalores de las 30 observaciones conocidas.

Se ha probado del paso 3 y del paso 6 que las componentes principales usando la matriz de covarianzas no han sido alterados respecto a su valor original con todos los datos conocidos y con datos faltantes estimados mediante las medias.

4.3 Análisis de Componentes para 4 variables

CASO 2 A Análisis Comparativo con 4 variables usando la Matriz de Correlaciones

Parte 1: <u>Con 30 observaciones todas conocidas</u>

Paso 1: La matriz de datos X_{30x4}

(Ver anexo)

Paso 2. La matriz de correlaciones para 4 variables para la muestra de tamaño 30

Es

$$R_n = \begin{bmatrix} 1 & -0.2982910 & 0.052411692 & -0.240816531 \\ -0.2982910 & 1 & 0.298747651 & 0.228681096 \\ 0.05241169 & 0.2987477 & 1 & -0.004705703 \\ -0.24081653 & 0.2286811 & -0.004705703 & 1 \end{bmatrix}$$

Paso 3 Identificación de los elementos del análisis de componentes principales, es decir definir la ecuación de la primera componente principal

$$Y_1 = 0.5247\left(\frac{X_1 - 0.5403}{\sqrt{0.0109068}}\right) - 0.6363\left(\frac{X_2 - 0.516}{\sqrt{0.0100091}}\right)$$
$$- 0.2882\left(\frac{X_3 - 0.547}{\sqrt{0.0142217}}\right) - 0.4864\left(\frac{X_4 - 0.5283}{\sqrt{0.0079385}}\right)$$

$$Y_1 = 5.02414\, X_1 - 6.36\, X_2 - 2.4167\, X_3 - 5.4591\, X_4 - 1.79028$$

La segunda componente principal viene a estar dado por

$$Y_2 = -0.41438\left(\frac{X_1 - 0.5403}{\sqrt{0.0109068}}\right) - 0.2557\left(\frac{X_2 - 0.516}{\sqrt{0.010091}}\right)$$
$$- 0.7961\left(\frac{X_3 - 0.547}{\sqrt{0.0142217}}\right) + 0.35928\left(\frac{X_4 - 0.5283}{\sqrt{0.0079385}}\right)$$

$$Y_2 = -3.9678\, X_1 - 2.5454\, X_2 - 6.675624\, X_3 + 4.0324\, X_4 + 4.9785$$

Parte 2: Con 31 observaciones donde la 31 observación es perdida,

Paso 4. Considerar la mismos datos donde la 31 ava observación se estima con la media de las 30 observaciones conocidas, resultando la misma matriz de datos estimada las observaciones perdidas con la media.

Paso 5 La matriz de datos X_{31x4}

(Ver anexo)

Paso 6 La matriz de correlaciones, los autovalores y la matriz de auvectores se muestra a continuación de las 31 observaciones

$$R_{n\times 1} = \begin{bmatrix} 1 & -0.2914814 & 0.0521626 & -0.2422241 \\ -0.2914814 & 1 & 0.2928326 & 0.2277192 \\ 0.0521626 & 0.2928326 & 1 & -0.0010027 \\ -0.2422241 & 0.2277192 & -0.001002797 & 1 \end{bmatrix}$$

Resumen : Cuadro Comparativo Nº4

	Vector de Medias	*Autovalores*	*Matriz de Autovectores*
n=30	0,540	1,55614	0.5247666 -0.4143801 0.5809266 0.4641481
	0,516	1,12538	-0.6363233 -0.2557623 -0.1725410 0.7070417
	0,547	0,77174	-0.2882481 -0.7961085 0.1086807 -0.5208769
	0,5283	0,54673	-0.4864419 0.3592854 0.7879990 -0.1155244
n=31	0,540	1,55614	0.5247666 -0.4143801 0.5809266 0.4641481
	0,516	1,12538	-0.6363233 -0.2557623 -0.1725410 0.7070417
	0,547	0,77174	-0.2882481 -0.7961085 0.1086807 -0.5208769
	0,5283	0,54673	-0.4864419 0.3592854 0.7879990 -0.1155244

Paso 7 Identificación de los elementos del análisis de componentes principales, es decir definir la ecuación de la primera componente principal

$$Y_1 = 0.5247\left(\frac{X_1 - 0.5403}{\sqrt{\left(\frac{29}{30}\right)0.0109068}}\right) - 0.6363\left(\frac{X_2 - 0.516}{\sqrt{\left(\frac{29}{30}\right)0.010091}}\right)$$
$$- 0.2882\left(\frac{X_3 - 0.547}{\sqrt{\left(\frac{29}{30}\right)0.0142217}}\right) - 0.4864\left(\frac{X_4 - 0.5283}{\sqrt{\left(\frac{29}{30}\right)0.0079385}}\right)$$

$$Y_1 = 5.11X_1 - 6.4425\,X_2 - 2.45798\,X_3 - 5.5525\,X_4 + 3.510592$$

La segunda componente principal viene a estar dado por

$$Y_2 = -0.4143\left(\frac{X_1 - 0.5403}{\sqrt{\left(\frac{29}{30}\right)0.0109068}}\right) - 0.2557\left(\frac{X_2 - 0.5162}{\sqrt{\left(\frac{29}{30}\right)0.010091}}\right)$$
$$- 0.7961\left(\frac{X_3 - 0.5473}{\sqrt{\left(\frac{29}{30}\right)0.0142217}}\right) + 0.3592\left(\frac{X_4 - 0.5283}{\sqrt{\left(\frac{29}{30}\right)0.0079385}}\right)$$

$$Y_1 = -0.403485\,X_1 - 2.5889\,X_2 - 6.789745\,X_3 + 4.1004\,X_4 + 9.39865193$$

Paso 8

-Comparando resultados del paso 3 y paso 7 se obtiene que los elementos del análisis de componentes principales usando la matriz de correlaciones $\mathbf{R}_{(n)}$ y al ser comparado con los elementos del análisis componentes principales usando la matriz

de correlaciones $\mathbf{R}_{(n+1)}$, son iguales las componentes con las 30 observaciones conocidas.

CASO 2B: Análisis comparativo con 4 variables usando la matriz de covarianzas sesgadas

Parte 1: <u>Con 30 observaciones todas conocidas</u>

Paso 1 La matriz de datos X_{30x4}

(Ver anexo)

Paso 2. La matriz de covarianzas sesgada para 4 variables para la muestra de tamaño 30

$$S_n = \begin{bmatrix} 0.010543 & -0.003012 & -0.000631 & -0.02166 \\ -0.003012 & 0.009670 & 0.003444 & 0.00197 \\ 0.000631 & 0.003444 & 0.013747 & -0.000048 \\ -0.002166 & 0.00197 & -0.0000483 & 0.0076738 \end{bmatrix}$$

Parte 2: Con 31 observaciones donde la 31 observación es perdida

Paso 3. Considerar los mismos datos donde la 31 ava observación se estima

con la media de las 30 observaciones conocidas, resultando la misma

Matriz de datos estimada las observaciones perdidas con la media

Paso 4 Obtener la matriz de covarianzas sesgada,los autovalores y la matriz de autovectores de la matriz con 31 datos

$$S_{n+1} = \begin{bmatrix} 0.01019 & -0.0029 & 0.00061 & -0.00209 \\ -0.0029 & 0.00934 & 0.00333 & 0.00190 \\ 0.00061 & 0.00333 & 0.01328 & -0.000046 \\ -0.00209 & 0.00190 & -0.00004 & 0.007418 \end{bmatrix}$$

Resumen : Cuadro Comparativo Nº5

	Vector de Medias	*Autovalores*	*Matriz de Autovectores*
n=30	0.540	0.01624	$\begin{bmatrix} 0.307 & 0.694 & -0.549 & -0.35 \\ -0.582 & -0.228 & 0.122 & -0.771 \\ -0.724 & 0.581 & -0.10 & 0.359 \\ -0.207 & -0.36 & 0.821 & 0.392 \end{bmatrix}$
	0.516	0.01318	
	0.547	0.00652	
	0.5283	0.00569	
n=31	0.540	0.01570	$\begin{bmatrix} 0.307 & 0.694 & -0.549 & -0.35 \\ -0.582 & -0.228 & 0.122 & -0.771 \\ -0.724 & 0.581 & -0.10 & 0.359 \\ -0.207 & -0.36 & 0.821 & 0.392 \end{bmatrix}$
	0.516	0.01274	
	0.547	0.00631	
	0.5283	0.00550	

Paso 5

Comparar los resultados del paso 2 con los del paso 4 se demuestra que la

la varianza muestral sesgada al tener la 31ava observación perdida queda afectado por el factor $(\frac{n}{n+1})$ de la varianza muestral sesgada al tener 30 observaciones conocidas.

Al observar el cuadro comparativo Nº5, los autovectores al tener 30 observaciones conocidas son iguales al tener la 31ava observación perdida estimada por la media de las 30 observaciones conocidas, en cambio con las 31 observaciones los autovalores

al tener la 31ava observación perdida estimada por la media de las 30 observaciones conocidas, disminuye en $(\frac{n}{n+1})$ de los autovalores de las 30 observaciones conocidas.

CASO 2C

Análisis Comparativo con 4 variables usando la Matriz de Covarianzas Insesgada

Parte 1: <u>Con 30 observaciones todas conocidas</u>

La matriz de datos X_{30x4}

Paso 1. La matriz de covarianzas insesgada para 4 variables para la muestra de tamaño 30 es

$$V_{n} = \begin{bmatrix} 0.0109068 & -0.00311586 & 0.0006527 & -0.0022408 \\ -0.0031158 & 0.0100041 & 0.0035634 & 0.0020379 \\ 0.0006527 & 0.003563 & 0.0142217 & -0.00005 \\ -0.00224 & 0.002037 & -0.00005 & 0.0079385 \end{bmatrix},$$

Paso 2 Identificación de los elementos del análisis de componentes principales, es decir definir la ecuación de la primera componente principal

$$Y_1 = 0.3067X_1 - 0.5824X_2 - 0.7237X_3 - 0.2075X_4$$

La segunda componente principal viene a estar dada por:

$$Y_2 = 0.694X_1 - 0.2276X_2 + 0.5805X_3 - 0.3595X_4$$

Parte 2: Con 31 observaciones donde la 31 observación es perdida

Paso 3. Considerar la mismos datos donde la 31 ava observación se estima con la media de las 30 observaciones conocidas, resultando la misma matriz de datos estimada las observaciones perdidas con la media

La matriz de datos X_{31x4}

(Ver anexo)

Paso 4 La matriz de covarianzas de la matriz con 31 datos es

$$V_{n+1} = \begin{bmatrix} 0.01054 & -0.003012 & 0.0006310 & -0.002166 \\ -0.003012 & 0.009670 & 0.0044667 & 0.001970 \\ 0.0006310 & 0.00344667 & 0.01374767 & -0.00004833 \\ -0.002166 & 0.001970 & -0.00004833 & 0.007673889 \end{bmatrix}$$

Paso 5 Identificación de los elementos del análisis de componentes principales, es decir definir la ecuación de la primera y segunda componente principal

$$Y_1 = 0.3067X_1 - 0.5824X_2 - 0.7237X_3 - 0.2075X_4$$

La segunda componente principal viene a estar dada por:

$$Y_2 = 0.694X_1 - 0.2276X_2 + 0.5805X_3 - 0.3595X_4$$

Paso 6 Del cuadro comparativo Nº6

Resumen : Cuadro Comparativo Nº6

	Vector de Medias	*Autovalores*	*Matriz de Autovectores*
n=30	0,540	0,01680	$\theta = \begin{bmatrix} 0.3067285 & 0.6941381 & 0.5490532 & -0.3501863 \\ -0.5824123 & -0.2276953 & 0.1215975 & -0.7708209 \\ -0.72365559 & 0.5805493 & -0.1004350 & 0.3594413 \\ -0.2074518 & -0.3595688 & 0.8207725 & 0.3924368 \end{bmatrix}$
	0,516	0,01364	
	0,547	0,00675	
	0,5283	0,00589	
n=31	0,540	0,01624	$\delta = \begin{bmatrix} 0.3067285 & 0.6941381 & 0.5490532 & -0.3501863 \\ -0.5824123 & -0.2276953 & 0.1215975 & -0.7708209 \\ -0.72365559 & 0.5805493 & -0.1004350 & 0.3594413 \\ -0.2074518 & -0.3595688 & 0.8207725 & 0.3924368 \end{bmatrix}$
	0,516	0,01318	
	0,547	0,00652	
	0,5283	0,00569	

Se demuestra que los autovectores teniendo los 30 datos conocidos

es igual a los autovectores con 31 observaciones donde la 31 observación

es estimada por la media de las 30 observaciones conocidas.

$$\theta_{1(n)} = \delta_{1(n+1)}, \theta_{2(n)} = \delta_{2(n+1)}, \theta_{3(n)} = \delta_{3(n+1)}., \theta_{4(n)} = \delta_{4(n+1)}$$

-Se demuestra que

$$\gamma_{1(n+1)=}\left(\frac{29}{30}\right)\lambda_{1(n)}, \gamma_{2(n+1)=}\left(\frac{29}{30}\right)\lambda_{2(n)}, \gamma_{3(n+1)=}\left(\frac{29}{30}\right)\lambda_{3(n)}, \gamma_{4(n+1)=}\left(\frac{29}{30}\right)\lambda_{4(n)}$$

Del cuadro comparativo Nº6 se demuestra que los autovalores, teniendo 31 observaciones, con la 31 ava observación perdida estimada por la media de las 30 observaciones conocidas, disminuyen en un

$\left(\frac{n-1}{n}\right)$ de los autovalores teniendo las 30 observaciones conocidas.

Los autovectores es invariante al número de observaciones.

Se ha probado del paso 2 y del paso 5 que las componentes principales usando la matriz de covarianzas no han sido alterados respecto a su valor original con todos los datos conocidos y con datos faltantes estimados mediante las medias.

4.4 Análisis de Componentes Principales para 5 variables

CASO 3A Análisis Comparativo con 5 variables usando la Matriz de Correlaciones

Parte 1: <u>Con 30 observaciones todas conocidas</u>

Paso 1 La matriz de datos X_{30x5}

Paso 2. La matriz de correlaciones para 5 variables

para la muestra de tamaño 30 es

$$R_{n} = \begin{bmatrix} 1 & 0.175484531 & -0.2224700 & 0.262903854 & 0.13807984 \\ 0.1754845 & 1 & -0.1730206 & 0.008451879 & -0.01281765 \\ -0.2224700 & -0.173020629 & 1 & -0.207125724 & -0.27076298 \\ 0.2629039 & 0.008451879 & -0.2071257 & 1 & 0.32120600 \\ 0.1380798 & -0.012817647 & -0.2707630 & 0.321205995 & 1 \end{bmatrix}$$

Paso 3 Identificación de los elementos del análisis de componentes principales,

es definir la ecuación de la primera componente principal

$$Y_1 = -0.4659\left(\frac{X_1 - 0.5532}{\sqrt{0.014601034}}\right) - 0.2225\left(\frac{X_2 - 0.5713}{\sqrt{0.01140644}}\right) + 0.5032\left(\frac{X_3 - 0.5524}{\sqrt{0.028478621}}\right) - 0.5030\left(\frac{X_4 - 0.52454}{\sqrt{0.01445161}}\right) - 0.4765\left(\frac{X_5 - 0.5142}{\sqrt{0.0144874713}}\right)$$

$$Y_1 = -3.855679\, X_1 - 2.083315\, X_2 + 2.9818\, X_3 + 4.184176\, X_4 - 3.95882\, X_5 + 5.9033854$$

La segunda componente principal viene a ser:

$$Y_2 = 0.2560\left(\frac{X_1 - 0.5532}{\sqrt{0.014601034}}\right) + 0.7890\left(\frac{X_2 - 0.5713}{\sqrt{0.01140644}}\right) - 0.1479\left(\frac{X_3 - 0.5524}{\sqrt{0.028478621}}\right) -$$
$$0.3327\left(\frac{X_4 - 0.52454}{\sqrt{0.01445161}}\right) - 0.4237\left(\frac{X_5 - 0.5142}{\sqrt{0.0144874713}}\right)$$

6 $$Y_2 = 2.118596\,X_1 + 7.387576\,X_2 - 0.876412\,X_3 - 2.76754\,X_4 - 3.52016\,X_5 - 1.636646$$

Parte 2: Con 31 observaciones donde la 31 observación es perdida

7 **Paso 4**. Considerar la mismos datos donde la 31 ava observación se estima con la media de las 30 observaciones conocidas, resultando la misma matriz de datos estimada las observaciones perdidas con la media

La matriz de datos X_{31x5}

Paso 5 La matriz de correlaciones para los 31 datos es

$$R_{n+1} = \begin{bmatrix} 1 & 0.175484531 & -0.2224700 & 0.262903854 & 0.13807984 \\ 0.1754845 & 1 & -0.1730206 & 0.008451879 & -0.01281765 \\ -0.2224700 & -0.173020629 & 1 & -0.207125724 & -0.27076298 \\ 0.2629039 & 0.008451879 & -0.2071257 & 1 & 0.32120600 \\ 0.1380798 & -0.012817647 & -0.2707630 & 0.321205995 & 1 \end{bmatrix}$$

Paso 6 Identificación de los elementos del análisis de componentes principales, es definir la ecuación de la primera componente principal

$$Y_1 = -0.4659\left(\frac{X_1 - 0.5532}{\sqrt{\left(\frac{29}{30}\right)0.014601034}}\right) - 0.2225\left(\frac{X_2 - 0.5713}{\sqrt{\left(\frac{29}{30}\right)0.011406}}\right) + 0.5032\left(\frac{X_3 - 0.5524}{\sqrt{\left(\frac{29}{30}\right)0.0284786}}\right) -$$
$$0.5030\left(\frac{X_4 - 0.52454}{\sqrt{\left(\frac{29}{30}\right)0.01445161}}\right) - 0.4765\left(\frac{X_5 - 0.5142_5}{\sqrt{\left(\frac{29}{30}\right)0.0144874713}}\right)$$

$$Y_1 = -3.921593\,X_1 - 2.118970\,X_2 + 3.03279\,X_3 + 4.2557\,X_4 - 4.02650\,X_5 + 1.542819$$

La segunda componente principal

$$Y_2 = 0.2560\left(\frac{X_1 - 0.5532}{\sqrt{\left(\frac{29}{30}\right)0.014601034}}\right) + 0.7890\left(\frac{X_2 - 0.5713}{\sqrt{\left(\frac{29}{30}\right)0.01140644}}\right) -$$

$$0.1479\left(\frac{X_3 - 0.5524}{\sqrt{\left(\frac{29}{30}\right)0.028478621}}\right) - 0.3327\left(\frac{X_4 - 0.52454}{\sqrt{\left(\frac{29}{30}\right)0.01445161}}\right)$$

$$-0.4237\left(\frac{X_5 - 0.5142}{\sqrt{\left(\frac{29}{30}\right)0.0144874713}}\right)$$

$$Y_2 = 2.154814\,X_1 + 7.513868\,X_2 + 0.891395\,X_3 - 2.814485\,X_4 - 3.580336\,X_5 - 5.134952$$

Paso 7

Resumen : Cuadro Comparativo Nº7

	Vector de Medias	*Autovalores*	*Matriz de Autovectores*				
n=30	0,553	1,75259	-0.4659	0.2560	0.6507	-0.3920	0.3743
	0,5713	1,10105	-0.2225	0.7890	-0.1587	0.5493	0.03474
	0,5524	0,83732	0.5032	-0.1479	0.5378	0.5485	0.3670
	0,5246	0,68318	-0.5030	-0.3327	0.3549	0.4135	-0.5825
	0,5142	0,62580	-0.4765	-0.4237	-0.3688	0.2696	0.6201
n=31	0,553	1,75259	-0.4659	0.2560	0.6507	-0.3920	0.3743
	0,5713	1,10105	-0.2225	0.7890	-0.1587	0.5493	0.03474
	0,5524	0,83732	0.5032	-0.1479	0.5378	0.5485	0.3670
	0,5246	0,68318	-0.5030	-0.3327	0.3549	0.4135	-0.5825
	0,5142	0,62580	-0.4765	-0.4237	-0.3688	0.2696	0.6201

Comparando resultados del paso 3 y paso 6 se obtiene que los elementos del análisis de componentes principales usando la matriz de correlaciones $\mathbf{R}_{(n)}$ y al ser comparado con los elementos del análisis componentes principales usando la matriz de correlaciones $\mathbf{R}_{(n+1)}$, son casi exactamente iguales las componentes con las 31 observaciones.

CASO 3B: Análisis comparativo con 5 variables usando la matriz de covarianzas sesgadas

Parte 1: <u>Con 30 observaciones todas conocidas</u>

Paso 1 La matriz de datos X_{30x5}

(Ver anexo)

Paso 2. La matriz de covarianzas sesgadas para 5 variables para la muestra de tamaño 30 con todas las observaciones conocidas es

$$S_{n} = \begin{bmatrix} 0.014114333 & 0.00222800 & -0.004346000 & 0.003695667 & 0.001993667 \\ 0.002228000 & 0.01102622 & -0.003178000 & 0.000030889 & -0.0002295556 \\ -0.004346000 & -0.00317800 & 0.027529333 & -0.004087333 & -0.005378667 \\ 0.003695667 & 0.000030889 & -0.004087333 & 0.01396989 & 0.0044174444 \\ 0.001993667 & -0.0002295556 & -0.005378667 & 0.0044174444 & 0.0140045556 \end{bmatrix}$$

Parte 2: Con 31 observaciones donde la 31 observación es perdida

Paso 3. Considerar la mismos datos donde la 31 ava observación se estima con la media de las 30 observaciones conocidas, resultando la misma matriz de datos estimada las observaciones perdidas con la media.

La matriz de datos X_{31x5}

Paso 4 Obtener la matriz de covarianzas, autovalores, y la matriz de autovectores con 31 datos

$$S_{n+1} = \begin{bmatrix} 0.01364385 & 0.00215373 & -0.004201133 & 0.003572478 & 0.001927211 \\ 0.002153733 & 0.01065868 & -0.003072067 & 0.00002985 & -0.0002219037 \\ -0.00420113 & -0.0030720 & 0.026611689 & -0.003951089 & -0.005199378 \\ 0.003572478 & 0.00002985 & -0.003951089 & 0.013500 & 0.0042701963 \\ 0.001927211 & -0.0002219 & -0.005199378 & 0.0042701963 & 0.0135377370 \end{bmatrix}$$

Resumen : Cuadro Comparativo Nº8

	Vector de Medias	*Autovalores*	*Matriz de Autovectores*
n=30	0,553 0,5713 0,5524 0,5246 0,5142	0,03446 0,01639 0,01380 0,00970 0,00907	$\begin{bmatrix} -0.2979454 & -0.3489458 & 0.70837517 & -0.1343396 & 0.51925221 \\ -0.1449710 & 0.1869362 & 0.51988506 & 0.6525117 & -0.49798221 \\ 0.8285225 & -0.5024809 & 0.15035825 & 0.1953767 & -0.01684603 \\ -0.3076979 & -0.6290019 & -0.04169476 & -0.3350178 & -0.62904951 \\ -0.3302793 & -0.4417706 & -0.45119629 & 0.6370027 & 0.29394521 \end{bmatrix}$
n=31	0,553 0,5713 0,5524 0,5246 0,5142	0,03220 0,01532 0,01289 0,00906 0,00848	$\begin{bmatrix} -0.2979454 & -0.3489458 & 0.70837517 & -0.1343396 & 0.51925221 \\ -0.1449710 & 0.1869362 & 0.51988506 & 0.6525117 & -0.49798221 \\ 0.8285225 & -0.5024809 & 0.15035825 & 0.1953767 & -0.01684603 \\ -0.3076979 & -0.6290019 & -0.04169476 & -0.3350178 & -0.62904951 \\ -0.3302793 & -0.4417706 & -0.45119629 & 0.6370027 & 0.29394521 \end{bmatrix}$

Paso 5

Se demuestra que la varianza muestral sesgada con 31 observaciones, al tener la 31ava observación perdida estimada por la media de las 30 observaciones conocidas queda afectado por un factor $(\frac{n}{n+1})$ de la varianza muestral sesgada al tener las 30 observaciones conocidas y los autovalores teniendo la 31ava observación perdida disminuyen en $(\frac{n}{n+1})$ en relación de los autovalores con todos los 30 datos conocidos.

Del cuadro comparativo Nº 8 se demuestra que los autovalores, teniendo 31 observaciones, con la 31 ava observación perdida estimada por la media de las 30 observaciones conocidas disminuyen en $(\frac{n}{n+1})$ en comparación con los 30 datos conocidos ,los autovectores se mantienen.

CASO 3C Análisis Comparativo con 5 variables usando la Matriz de Covarianzas Insesgada

Parte 1: <u>Con 30 observaciones todas conocidas</u>

Paso 1 La matriz de datos X_{30x5}

(Ver anexo)

Estos datos han sido simulados con distribución normal estándar.

Paso 2. La matriz de covarianzas para 5 variables para la muestra de tamaño 30 es

$$V_x = \begin{bmatrix} 0.014601034 & 0.002304828 & -0.004495862 & -0.004495862 & 0.002062414 \\ 0.002304828 & 0.01140644 & -0.003287586 & 0.00003195402 & -0.0002374713 \\ -0.004495862 & -0.003287586 & 0.028478621 & -0.004228276 & -0.005564138 \\ 0.03823103 & 0.00003195402 & -0.004228276 & 0.01445161 & 0.0045697701 \\ 0.002062414 & -0.0002374713 & -0.005564138 & 0.004569770 & 0.0144874713 \end{bmatrix}$$

Paso 3 Identificación de los elementos del análisis de componentes principales, es decir definir la ecuación de la primera componente principal

$$Y_1 = -0.298 X_1 - 0.145 X_2 + 0.829 X_3 - 0.308 X_4 - 0.33 X_5$$

Parte 2: Con 31 observaciones donde la 31 observación es perdida

Paso 4. Considerar la mismos datos donde la 31 ava observación se estima con la media de las 30 observaciones conocidas, resultando la misma matriz de datos estimada las observaciones perdidas con la media.

Paso 5 Obtener la matriz de covarianzas de la matriz con 31 datos

$$V_{n+1} = \begin{bmatrix} 0.014114333 & 0.002228000 & -0.004346000 & 0.003695667 & 0.001993667 \\ 0.002228000 & 0.0102622 & -0.003178000 & 0.00003088889 & -0.0002295556 \\ -0.004346000 & -0.003178000 & 0.027529333 & -0.004087333 & -0.005378667 \\ 0.003695667 & 0.00003088889 & -0.004087333 & 0.01396989 & 0.0044174444 \\ 0.001993667 & -0.0002295556 & -0.005378667 & 0.004417444 & 0.0140045556 \end{bmatrix}$$

Paso 6

Resumen : Cuadro Comparativo Nº9

	Vector de Medias	*Autovalores*	*Matriz de Autovectores*
n=30	0,553	0,03446	$\theta =$ [-0.2979454 -0.3489458 0.70837517 -0.1343396 0.5192522; -0.1449710 0.1869362 0.51988506 0.6525117 -0.4979822; 0.8285225 -0.5024809 0.15035825 0.1953767 -0.0168460; -0.3076979 -0.6290019 -0.04169476 -0.3350178 -0.6290495; -0.3302793 -0.4417706 -0.45119629 0.6370027 0.2939452]
	0,5713	0,01639	
	0,5524	0,01380	
	0,5246	0,00970	
	0,5142	0,00907	
n=31	0,553	0,03333	$\delta =$ [-0.2979454 -0.3489458 0.70837517 -0.1343396 0.5192522; -0.1449710 0.1869362 0.51988506 0.6525117 -0.4979822; 0.8285225 -0.5024809 0.15035825 0.1953767 -0.0168460; -0.3076979 -0.6290019 -0.04169476 -0.3350178 -0.6290495; -0.3302793 -0.4417706 -0.45119629 0.6370027 0.2939452]
	0,5713	0,01585	
	0,5524	0,01334	
	0,5246	0,00938	
	0,5142	0,00877	

Al comparar los autovalores teniendo 31 observaciones donde la 31 observación es estimada por las 30 observaciones conocidas disminuye en $(n-1/n)$ de los autovalores teniendo los 30 datos conocidos. Los autovectores es invariante al número de observaciones.

Identificación de los elementos del análisis de componentes principales, es decir definir la ecuación de la primera componente principal

$$Y_1 = -0.298\,X_1 - 0.145X_2 + 0.829\,X_3 - 0.308\,X_4 - 0.33\,X_5$$

Paso 7

Del cuadro comparativo:

Se demuestra que los autovectores teniendo los 30 datos conocidos

es igual a los autovectores con 31 observaciones donde la 31 observación es perdida

$$\theta_{1(n)} = \delta_{1(n+1)}, \quad \theta_{2(n)} = \delta_{2(n+1)}, \theta_{3(n)} = \delta_{3(n+1)}.$$

$$\theta_{4(n)} = \delta_{4(n+1)}, \theta_{5(n)} = \delta_{5(n+1)}$$

-Se demuestra que

$$\gamma_{1(n+1)=}\left(\frac{29}{30}\right)\lambda_{1(n)}, \gamma_{2(n+1)=}\left(\frac{29}{30}\right)\lambda_{2(n)}, \gamma_{3(n+1)=}\left(\frac{29}{30}\right)\lambda_{3(n)}, \gamma_{4(n+1)=}\left(\frac{29}{30}\right)\lambda_{4(n)}, \gamma_{5(n+1)=}\left(\frac{29}{30}\right)\lambda_{5(n)}$$

Por lo que al comparar con el paso 3 se demuestra que los autovalores con 31 observaciones con la 31 ava observación perdida estimada por la media de las 30 observaciones conocidas, disminuyen en $\left(\frac{n-1}{n}\right)$ de los autovalores de las 30 observaciones conocidas.

Se ha probado del paso 3 y del paso 6 que las componentes principales usando la matriz de covarianzas no han sido alterados respecto a su valor original con todos los datos conocidos y con datos faltantes estimados mediante las medias.

4.5 Análisis de Componentes Principales para 6 variables

CASO 4A: Análisis comparativo con 6 variables usando la matriz de correlaciones

Parte 1: <u>Con 30 observaciones todas conocidas</u>

Paso 1 La matriz de datos X_{30x6}

Paso 2. La matriz de correlación para 6 variables para la muestra de tamaño 30 con todas las observaciones conocidas es

$$R_n = \begin{bmatrix} 1 & -0.1310251 & 0.060279901 & -0.06703644 & 0.30948581 & 0.237700561 \\ -0.13102507 & 1 & -0.126531811 & -0.09334579 & -0.30561627 & -0.172664301 \\ 0.06027990 & -0.1265318 & 1 & -0.15489151 & 0.27217304 & 0.007150916 \\ -0.06703644 & -0.0933458 & -0.154891514 & 1 & -0.01287612 & -0.219504887 \\ 0.30948581 & -0.3056163 & 0.272173038 & -0.01287612 & 1 & 0.268538732 \\ 0.23770056 & -0.1726643 & 0.007150916 & -0.21950489 & 0.26853873 & 1 \end{bmatrix},$$

Paso 3 Identificación de los elementos del análisis de componentes principales, es decir definir la ecuación de la primera componente principal.

$$Y_1 = 0.44007\left(\frac{X_1 - 0.546}{\sqrt{0.01974}}\right) - 0.3999517\left(\frac{X_2 - 0.5697}{\sqrt{0.01175}}\right) +$$
$$0.3163780\left(\frac{X_3 - 0.516}{\sqrt{0.008631724}}\right) - 0.1765671\left(\frac{X_4 - 0.548}{\sqrt{0.0166}}\right)$$
$$+ 0.5642658\left(\frac{X_5 - 0.5273}{\sqrt{0.0142960}}\right) + 0.4435163\left(\frac{X_6 - 0.4973}{\sqrt{0.02234437}}\right)$$

$$Y_1 = -2.457659\,X_1 - 3.6857\,X_2 + 3.4053\,X_3 - 1.370424\,X_4 + 4.719287\,X_5 + 2.967054\,X_6 - 4.571176$$

La segunda componente principal es

$$Y_2 = 0.020732\left(\frac{X_1 - 0.546}{\sqrt{0.01974}}\right) + 0.48035027\left(\frac{X_2 - 0.5697}{\sqrt{0.011175747}}\right) +$$
$$0.007852585\left(\frac{X_3 - 0.516}{\sqrt{0.008631724}}\right) - 0.79808676\left(\frac{X_4 - 0.548}{\sqrt{0.0166644}}\right) +$$
$$0.19973435\left(\frac{X_5 - 0.5273}{\sqrt{0.01429609}}\right) + 0.29296942\left(\frac{X_6 - 0.4973}{\sqrt{0.02234437}}\right)$$

$$Y_2 = 0.14756\,X_1 + 4.543805\,X_2 + 0.0845\,X_3 - 6.1823\,X_4 + 1.670490\,X_5 + 1.959919\,X_6 - 1.1803378369$$

La tercera componente principal es.

$$Y_3 = -0.345299\left(\frac{X_1 - 0.546}{\sqrt{0.01974}}\right) - 0.022852\left(\frac{X_2 - 0.5697}{\sqrt{0.011175}}\right) +$$
$$0.810219\left(\frac{X_3 - 0.516}{\sqrt{0.0086317}}\right) - 0.13159\left(\frac{X_4 - 0.548}{\sqrt{0.0166441379}}\right) +$$
$$0.10511344\left(\frac{X_5 - 0.5273}{\sqrt{0.014296092}}\right) - 0.442069\left(\frac{X_6 - 0.4973}{\sqrt{0.02234437}}\right)$$

$$Y_3 = -2.457659\ X_1 - 0.216172\ X_2 + 8.72\,X_3 - 1.0199\ X_4 + 0.8791\ X_5 - 2.9573\ X_6 - 2.93947$$

Parte 2: Con 31 observaciones donde la 31 observación es perdida

Paso 4. Considerar la mismos datos donde la 31 ava observación se estima con la media de las 30 observaciones conocidas, resultando la misma matriz de datos estimada las observaciones perdidas con la media.

Paso 5 Obtener la matriz de correlación, los autovalores y autovectores de la matriz con 31 datos.

$$R_{n+1} = \begin{bmatrix} 1 & -0.1310251 & 0.060279901 & -0.06703644 & 0.30948581 & 0.237700561 \\ -0.13102507 & 1 & -0.126531811 & -0.09334579 & -0.30561627 & -0.172664301 \\ 0.06027990 & -0.1265318 & 1 & -0.15489151 & 0.27217304 & 0.007150916 \\ -0.06703644 & -0.0933458 & -0.154891514 & 1 & -0.01287612 & -0.219504887 \\ 0.30948581 & -0.3056163 & 0.272173038 & -0.01287612 & 1 & 0.268538732 \\ 0.23770056 & -0.1726643 & 0.007150916 & -0.21950489 & 0.26853873 & 1 \end{bmatrix}$$

Resumen : Cuadro Comparativo Nº10

	Vector de Medias	*Autovalores*	*Matriz de Autovectores*					
n=30	0,546	1,8260	0.4400738	0.02073210	-0.3452942	0.6995037	0.39716032	-0.199077280
	0,5697	1,1510	-0.3999517	0.48035027	-0.02285252	0.5452988	-0.55799568	0.008369244
	0,516	1,0340	0.3163780	0.07852585	0.81021980	0.1153448	-0.08661033	-0.465270589
	0,548	0,8200	-0.1765671	-0.79808676	-0.13159019	0.1823167	-0.38243036	-0.367523084
	0,5273	0,6370	0.5642658	-0.19973435	0.10511344	0.1360244	-0.43803924	0.648290361
	0,4973	0,5330	0.4435163	0.29296962	-0.44206915	-0.3850939	-0.43043210	-0.434128854
n=31	0,546	1,8260	0.4400738	0.02073210	-0.3452942	0.6995037	0.39716032	-0.199077280
	0,5697	1,1510	-0.3999517	0.48035027	-0.02285252	0.5452988	-0.55799568	0.008369244
	0,516	1,0340	0.3163780	0.07852585	0.81021980	0.1153448	-0.08661033	-0.465270589
	0,548	0,8200	-0.1765671	-0.79808676	-0.13159019	0.1823167	-0.38243036	-0.367523084
	0,5273	0,6370	0.5642658	-0.19973435	0.10511344	0.1360244	-0.43803924	0.648290361
	0,4973	0,5330	0.4435163	0.29296962	-0.44206915	-0.3850939	-0.43043210	-0.434128854

Paso 6 Identificación de los elementos del análisis de componentes principales, es definir la ecuación de la primera componente principal.

$$Y_1 = 0.44007\left(\frac{X_1-0.546}{\sqrt{\left(\frac{29}{30}\right)0.01974}}\right) - 0.3999517\left(\frac{X_2-0.5697}{\sqrt{\left(\frac{29}{30}\right)0.011175747}}\right) +$$

$$0.3163780\left(\frac{X_3-0.516}{\sqrt{\left(\frac{29}{30}\right)0.008631724}}\right) - 0.1765671\left(\frac{X_4-0.548}{\sqrt{\left(\frac{29}{30}\right)0.0166441379}}\right)$$

$$+0.5642658\left(\frac{X_5-0.5273}{\sqrt{\left(\frac{29}{30}\right)0.0142960920}}\right) + 0.4435\left(\frac{X_6-0.4973}{\sqrt{\left(\frac{29}{30}\right)0.02234437}}\right)$$

$$Y_2 = 3.185736\,X_1 - 3.84838\,X_2 + 3.46353\,X_3 - 1.3920\,X_4 + 4.7999\,X_5 + 3.01766\,X_6 - 4.603278$$

.

La segunda componente principal viene a ser

$$Y_2 = 0.020732\left(\frac{X_1-0.546}{\sqrt{\left(\frac{29}{30}\right)0.01974}}\right) + 0.48035027\left(\frac{X_2-0.5697}{\sqrt{\left(\frac{29}{30}\right)0.011175}}\right) +$$

$$0.07852585\left(\frac{X_3-0.516}{\sqrt{\left(\frac{29}{30}\right)0.008631724}}\right) - 0.19973435\left(\frac{X_4-0.548}{\sqrt{\left(\frac{29}{30}\right)0.0166441379}}\right) +$$

$$0.19973435\left(\frac{X_5-0.5273}{\sqrt{\left(\frac{29}{30}\right)0.0142960920}}\right) + 0.292969\left(\frac{X_6-0.4973}{\sqrt{\left(\frac{29}{30}\right)0.02234437}}\right)$$

$$Y_2 = 0.15\,X_1 + 4.6216\,X_2 + 0.859657\,X_3 - 1.5746489\,X_4 + 1.6990\,X_5 + 1.19934\,X_6 - 4.182773816$$

.

La tercera componente principal viene a ser

$$Y_3 = -0.345299\left(\frac{X_1-0.546}{\sqrt{\left(\frac{29}{30}\right)0.01974}}\right) - 0.022852\left(\frac{X_2-0.5697}{\sqrt{\left(\frac{29}{30}\right)0.011175}}\right) +$$

$$0.810219\left(\frac{X_3-0.516}{\sqrt{\left(\frac{29}{30}\right)0.008631724}}\right) - 0.13159\left(\frac{X_4-0.548}{\sqrt{\left(\frac{29}{30}\right)0.0166441379}}\right) + 0.10511344\left(\frac{X_5-0.5273}{\sqrt{\left(\frac{29}{30}\right)0.014296092}}\right)$$

$$-0.442069\left(\frac{X_6-0.4973}{\sqrt{\left(\frac{29}{30}\right)0.02234437}}\right)$$

$$Y_3 = -2.4996\,X_1 - 0.21986\,X_2 + 8.869829\,X_3 - 1.0374\,X_4 + 0.8941517\,X_5 - 3.0079\,X_6 - 0.36655$$

Paso 9

-Comparando resultados del paso 3 y paso 6 se obtiene que los elementos del análisis de componentes principales usando la matriz de correlaciones $\mathbf{R}_{(n)}$ y al ser comparado con los elementos del análisis componentes principales usando la matriz de correlaciones $\mathbf{R}_{(n+1)}$, son casi exactamente iguales las componentes con las 30 observaciones conocidas.

CASO 4B: Análisis comparativo con 6 variables usando la matriz de covarianzas

Sesgadas

Parte 1: <u>Con 30 observaciones todas conocidas</u>

Paso 1 Señalar que es la misma matriz con 30 datos conocidos

La matriz de datos X_{30x6}

Estos datos han sido simulados con distribución normal estándar.

Paso 2. La matriz de covarianzas sesgada para 6 variables para la muestra de tamaño 30 con todas las observaciones conocidas es

$$S_n = \begin{bmatrix} 0.019084 & -0.00188 & 0.000760 & -0.001174 & 0.005026 & 0.004826 \\ -0.0018813 & 0.01080 & -0.001201 & -0.001230 & -0.0037342 & -0.002637 \\ 0.000760 & -0.00120 & 0.008344 & -0.001794 & 0.002922 & -0.002637 \\ -0.0011746 & -0.00123 & -0.001794 & 0.016089 & -0.000192 & -0.004092 \\ 0.005026 & -0.003734 & 0.002922 & -0.000192 & 0.013819 & 0.004639 \\ 0.004826 & -0.002637 & 0.000096 & -0.004092 & 0.004639 & 0.021599 \end{bmatrix}$$

Paso 3 La matriz de autovalores y autovectores de la matriz de covarianzas sesgadas:

Parte 2: Con 31 observaciones donde la 31 observación es perdida

Paso 4. Considerar la mismos datos donde la 31 ava observación se estima con la media de las 30 observaciones conocidas, resultando la misma matriz de datos estimada las observaciones perdidas con la media.

La matriz de datos X_{31x6}

(Ver anexo)

Paso 5 Consideremos la matriz de covarianzas sesgada con 31 observaciones donde la 31 observación es perdida

$$S_{n+1} = \begin{bmatrix} 0.0184478 & -0.0018186 & 0.0007353 & -0.001135 & 0.004858 & 0.004665 \\ -0.0018186 & 0.010443 & -0.001161 & -0.001189 & -0.003609 & -0.002549 \\ 0.0007353 & -0.001161 & 0.008065 & -0.001734 & 0.002825 & 0.000092 \\ -0.001135 & -0.001189 & -0.001734 & 0.01555 & -0.000185 & -0.003955 \\ 0.0048584 & -0.003609 & 0.0028252 & -0.000185 & 0.013358 & 0.004484 \\ 0.0046651 & -0.002549 & 0.0000928 & -0.003955 & 0.004484 & 0.020879 \end{bmatrix}$$

Paso 6 En el cuadro comparativo Nº11

Resumen : Cuadro Comparativo Nº11

	Vector de Medias	*Autovalores*	*Matriz de Autovectores*					
n=30	0,546	0,0305	0.5266	0.4265	0.6071	0.3896	– 0.1074	– 0.1000
	0,5697	0,0179	– 0.2091	– 0.2691	0.2979	0.3310	0.8269	– 0.0406
	0,516	0,0144	0.1052	– 0.0318	0.1128	– 0.5099	0.1596	– 0.8304
	0,548	0,0126	– 0.2324	0.7016	– 0.5003	0.3269	0.2076	– 0.2314
	0,5273	0,0080	0.4112	0.3259	0.0933	– 0.5232	0.4759	0.4646
	0,4973	0,0062	0.6670	– 0.3824	– 0.5205	0.3130	– 0.0978	– 0.1743
n=31	0,546	0,0296	0.5266	0.4265	0.6071	0.3896	– 0.1074	– 0.1000
	0,5697	0,0175	– 0.2091	– 0.2691	0.2979	0.3310	0.8269	– 0.0406
	0,516	0,0144	0.1052	– 0.0318	0.1128	– 0.5099	0.1596	– 0.8304
	0,548	0,0139	– 0.2324	0.7016	– 0.5003	0.3269	0.2076	– 0.2314
	0,5273	0,0077	0.4112	0.3259	0.0933	– 0.5232	0.4759	0.4646
	0,4973	0,0060	0.6670	– 0.3824	– 0.5205	0.3130	– 0.0978	– 0.1743

teniendo 31 observaciones donde la 31 observación perdida estimada por la media de las 30 observaciones conocidas,los autovalores disminuyen en $(n+1/n)$ al comparar los autovalores de las 30 observaciones conocidos.

Comparar los resultados del paso 3 con los del paso 6 se demuestra que la

la varianza muestral sesgada al tener la 31ava observación perdida estimada por la media de las 30 observaciones conocidas queda afectada por el factor

$(\frac{n}{n+1})$ de la varianza muestral sesgada al tener 30 observaciones conocidas.

CASO 4C: Análisis Comparativo con 6 variables usando la Matriz de Covarianzas Insesgada

Parte 1: <u>Con 30 observaciones todas conocidas</u>

Paso 1 Señalar que es la misma matriz de covarianzas insesgada con 30 datos conocidos.

La matriz de datos X_{30x6}

(Ver anexo)

Estos datos han sido simulados con distribución normal estándar.

Paso 2. La matriz de covarianzas para 6 variables para la muestra de tamaño 30 con todas las observaciones conocidas es

$$V_n = \begin{bmatrix} 0.0197420690 & -0.001946207 & 0.0007868966 & -0.0012151724 & 0.0051993103 & 0.004992414 \\ -0.0019462069 & 0.011175747 & -0.001242759 & -0.0012731034 & -0.0038629885 & -0.002728506 \\ 0.0007868966 & -0.001242759 & 0.008631724 & -0.0018565517 & 0.0030234483 & 0.00009931034 \\ -0.0012151724 & -0.001273103 & -0.001856552 & 0.0166441379 & -0.0001986207 & -0.004233103 \\ 0.0051993103 & -0.003862989 & 0.003023448 & -0.0001986207 & 0.0142960920 & 0.00479954 \\ 0.0049924138 & -0.002728506 & 0.00009931034 & -0.0042331034 & 0.0047995402 & 0.02234437 \end{bmatrix}$$

Paso 3 Identificación de los elementos del análisis de componentes principales, es decir definir la ecuación de la primera componente principal

$$Y_1 = 0.52660X_1 - 0.20905X_2 + 0.10518\ X_3 - 0.23243X_4 + 0.41115\ X_5 + 0.66696X_6$$

La segunda componente principal viene a ser:

$$Y_2 = 0.42647X_1 - 0.26912X_2 - 0.031818\ X_3 - 0.701569X_4\ - 0.32586X_5 - 0.38248\ X_6$$

Parte 2: Con 31 observaciones donde la 31 observación es perdida

Paso 4. Considerar la mismos datos donde la 31 ava observación se estima con la media de las 30 observaciones conocidas, resultando la misma matriz de datos estimada las observaciones perdidas con la media

La matriz de datos X_{31x6}

Paso 5 Obtener la matriz de covarianzas de la matriz con 31 datos

$$V_{n+1} = \begin{bmatrix} 0.019084 & -0.001881 & 0.00076 & -0.00117 & 0.005026 & 0.004826 \\ 0.001881 & 0.010803 & -0.00120 & -0.001230 & -0.003734 & -0.0026375 \\ 0.000760 & -0.00120 & 0.00834 & -0.001794 & 0.002922 & 0.000096 \\ -0.0011746 & -0.001230 & -0.00179 & 0.016089 & -0.000192 & -0.004092 \\ 0.005026 & -0.003734 & 0.002922 & -0.000192 & 0.013819 & 0.004639 \\ 0.004826 & -0.002637 & 0.000096 & -0.00409 & 0.004639 & 0.021599 \end{bmatrix}$$

Paso 6

Identificación de los elementos del análisis de componentes principales, es decir definir la ecuación de la primera componente principal

$$Y_1 = 0.52660X_1 - 0.20905X_2 + 0.10518\,X_3 - 0.23243X_4 + 0.41115\,X_5 + 0.66696X_6$$

La segunda componente principal viene a ser:

$$Y_2 = 0.42647X_1 - 0.26912X_2 - 0.031818\,X_3 - 0.701569X_4 - 0.32586X_5 - 0.38248\,X_6$$

Paso 7

En el cuadro comparativo que se muestra se tiene que

Resumen : Cuadro Comparativo Nº12

	Vector de Medias	*Autovalores*	*Matriz de Autovectores*
n=30	0,546 0,5697 0,516 0,548 0,5273 0,4973	0,0316 0,0185 0,1485 0,0131 0,0083 0,0065	$\theta = \begin{bmatrix} 0.52660 & 0.42647 & 0.607106 & 0.3896041 & -0.10740303 & -0.09437120 \\ -0.20905 & -0.26912 & 0.297957 & 0.3309933 & 0.82697295 & -0.04061574 \\ 0.10518 & 0.031818 & 0.112837 & -0.5099840 & 0.15962420 & -0.83043789 \\ -0.23243 & 0.701569 & -0.50026 & 0.3269183 & 0.20758909 & -0.23139772 \\ 0.41115 & 0.32586 & -0.09339 & -0.5231799 & 0.47585422 & 0.46463546 \\ 0.66696 & -0.38248 & -0.52050 & 0.3130105 & 0.09782625 & -0.17431679 \end{bmatrix}$
n=31	0,546 0,5697 0,516 0,548 0,5273 0,4973	0,0305 0,0179 0,0144 0,0126 0,0080 0,0063	$\delta = \begin{bmatrix} 0.52660 & 0.42647 & 0.607106 & 0.3896041 & -0.10740303 & -0.09437120 \\ -0.20905 & -0.26912 & 0.297957 & 0.3309933 & 0.82697295 & -0.04061574 \\ 0.10518 & 0.031818 & 0.112837 & -0.5099840 & 0.15962420 & -0.83043789 \\ -0.23243 & 0.701569 & -0.50026 & 0.3269183 & 0.20758909 & -0.23139772 \\ 0.41115 & 0.32586 & -0.09339 & -0.5231799 & 0.47585422 & 0.46463546 \\ 0.66696 & -0.38248 & -0.52050 & 0.3130105 & 0.09782625 & -0.17431679 \end{bmatrix}$

Del Cuadro comparativo Nº12 ,teniendo los 30 datos conocidos

es igual a los autovectores con 31 observaciones donde la 31 observación es perdida

$$\theta_{1(n)} = \delta_{1(n+1)}, \quad \theta_{2(n)} = \delta_{2(n+1)}, \theta_{3(n)} = \delta_{3(n+1)}.,$$

$$\theta_{4(n)} = \delta_{4(n+1)}, \theta_{5(n)} = \delta_{5(n+1)}, \theta_{6(n)} = \delta_{6(n+1)}$$

-Se demuestra que

$$\gamma_{1(n+1)=}\left(\frac{29}{30}\right)\lambda_{1(n)}, \gamma_{2(n+1)=}\left(\frac{29}{30}\right)\lambda_{2(n)}, \gamma_{3(n+1)=}\left(\frac{29}{30}\right)\lambda_{3(n)}, \gamma_{4(n+1)=}\left(\frac{29}{30}\right)\lambda_{4(n)},$$

$$\gamma_{5(n+1)=}\left(\frac{29}{30}\right)\lambda_{5(n)}, \gamma_{6(n+1)=}\left(\frac{29}{30}\right)\lambda_{6(n)}$$

Se demuestra que los autovalores con la 31 ava observación perdida que es estimada por la media de las 30 observaciones conocidas, disminuyen en un $\frac{(n-1)}{n}$

de los autovalores al tener 30 datos conocidos.Los autovectores son invariantes al número de observaciones.

Se ha probado del paso 3 y del paso 6 que las componentes principales usando la matriz de covarianzas no han sido alterados respecto a su valor original con todos los datos conocidos y con datos faltantes estimados mediante las medias.

CONCLUSIONES

Los principales resultados encontrados son:

1. Si se estiman datos faltantes mediante la media aritmética de las observaciones conocidas, la media aritmética de todos los datos, incluidos los faltantes, es igual a la media aritmética de los datos conocidos.

2. Al generalizar al caso multivariante, estimando separadamente para cada variable las observaciones perdidas mediante la media de las observaciones conocidas, el vector de medias con todos los datos incluidos los faltantes es igual al vector de medias con todas las observaciones conocidas.

3. La matriz de correlaciones con datos faltantes, no quedan afectadas por la estimación de dichos datos faltantes mediante la media aritmética de cada variable con datos conocidos.

4. La matriz de covarianzas sesgada con datos faltantes, quedan afectadas por la estimación de dichos datos mediante la media aritmética de cada variable por un factor $(n/n+1)$ de la matriz de covarianzas sesgada con todos los datos conocidos.

5. La matriz de covarianzas insesgada con datos faltantes, quedan afectadas por la estimación de dichos datos mediante la media aritmética de cada variable por un factor $(n-1/n)$ de la matriz de covarianzas insesgada con todos los datos conocidos.

6. Los autovalores y autovectores de las matrices de correlaciones, son los mismos para datos completos y con datos faltantes estimados mediante las medias.

7. Se ha probado que a través de la imputación de la media aritmética en donde se han encontrado algún dato faltante, los autovalores de la matriz de covarianzas V_{n+1} disminuyen aproximadamente en $(n-1/n)$ respecto de los autovalores de la matriz de covarianzas original V_n, los autovectores conservan su valor original.

8. Se ha probado que las componentes principales usando la matriz de covarianzas

   y correlaciones no han tenido cambios respecto a su valor original con todos los datos conocidos y con el dato imputado por la media aritmética.

REFERENCIAS BIBLIOGRÁFICAS

1. Anderson,T W. (1984) . *An Introduction to Multivariate Statistics Analysis.*(2 d).,Wiley e Sons,Inc,N.Y. pp 451-479.

2. Apostol, T. M.(1980) .*Cálculus. Vol 2.*(2 ed),España:Reverté.,1980. pp119-140.

3. Elong, L.L. (1998). *Algebra Lineal*.Imca.(2 ed),Perú:Hozlo.pp.316-326.

4. J, F.Hair. and J.R.E.Anderson.R.L and Tathawm.W.C.(1999) *Analysis Multivariate*,(5 ed),Madrid:Prentice Hall.,39-57.

5. Johnson,R.A. (2007).*Applied Multivariate Statistical Analysis*. (6 ed). London:Pearson.pp 430-465.

6. Mandeville, P.B. (2010),13(3).*Observaciones Perdidas.*México:Redalyc. Disponible en http://redalyc.uaemex.mx/redalyc/pdf/402/40215495020.pdf, pp 325-328.

7. Mardia, K.V.*" Multivariate Analysis*".(1979).London :Academia Press.pp 213-228.

8. Mood A.M and Graybill F.A.(1976.)."*Introducción a la Teoria Estadística*". (4 ed).Madrid: Aguilar.

9. Muñoz, R.J.(2007).*Cálculo diferencial&Integral.*Vol I.Brazil,Universidad Federal de Río de Janeiro.pp 159-220.

10. Ofer, H. and Shafer.(1987). *Multiple Imputation in two Stages.* University of Washington.Artículo Original publicado en Online.Disponible en http://www.fcsm.gov/03papers/Harel.pdf .

11. Olkin, I and Raveh , A. *(*2008). *Bounds for how much influence an observation can have.* Artículo Original publicado en Online. Disponible en http://www.fcsm.gov/03papers/Harel.pdf .

12. Rizzo,M. (2008).*Statistical Computing with R.*Chapman&Hall/CRC. Computer Science & Data Analysis.

13. Serge, L.Algebra Lineal.(1976).*(*2ed).México Fondo Educativo Interamericano.pp 167-275.

14. Sheldom, M. R.(1999). *Simulación.* (2 ed).México: Prentice Hall.pp 36.

15. Xiaowei, Y. and Shoptaw, S. *Multiple Partial Imputation for Longitudinal Data with Missing Values in Clinical Trials.* University of California. 2005.Artículo Original publicado en Online.Disponible en http://escholarship.org/uc/item/90b7b2xn;jsessionid=7E831BDD451B11635365ED45134E8EA6

ANEXOS

ANEXO A

DATOS QUE SE HAN OBTENIDO SIMULADO EN EL SOFTWARE LIBRE R

Anexo 8.a

- ✓ p=3, n=30

Teniendo las 30 observaciones

Conocidas

n	x1	x2	x3
1	0,56	0,53	0,67
2	0,55	0,35	0,41
3	0,66	0,32	0,65
4	0,58	0,56	0,44
5	0,38	0,53	0,65
6	0,51	0,53	0,53
7	0,6	0,44	0,45
8	0,3	0,5	0,56
9	0,64	0,52	0,7
10	0,69	0,49	0,64
11	0,57	0,51	0,5
12	0,52	0,32	0,55
13	0,39	0,49	0,5
14	0,47	0,62	0,41
15	0,53	0,46	0,58
16	0,41	0,49	0,45
17	0,63	0,84	0,62
18	0,65	0,27	0,61
19	0,61	0,62	0,41
20	0,46	0,58	0,38
21	0,55	0,73	0,52
22	0,45	0,76	0,68
23	0,52	0,79	0,59
24	0,31	0,36	0,69
25	0,41	0,42	0,58
26	0,73	0,6	0,38
27	0,48	0,42	0,48
28	0,82	0,39	0,59
29	0,72	0,47	0,58
30	0,58	0,68	0,67

Anexo 8.b

- ✓ p=3,n=31

n	x1	x2	x3
1	0,56	0,53	0,67
2	0,55	0,35	0,41
3	0,66	0,32	0,65
4	0,58	0,56	0,44
5	0,38	0,53	0,65
6	0,51	0,53	0,53
7	0,6	0,44	0,45
8	0,3	0,5	0,56
9	0,64	0,52	0,7
10	0,69	0,49	0,64
11	0,57	0,51	0,5
12	0,52	0,32	0,55
13	0,39	0,49	0,5
14	0,47	0,62	0,41
15	0,53	0,46	0,58
16	0,41	0,49	0,45
17	0,63	0,84	0,62
18	0,65	0,27	0,61
19	0,61	0,62	0,41
20	0,46	0,58	0,38
21	0,55	0,73	0,52
22	0,45	0,76	0,68
23	0,52	0,79	0,59
24	0,31	0,36	0,69
25	0,41	0,42	0,58
26	0,73	0,6	0,38
27	0,48	0,42	0,48
28	0,82	0,39	0,59
29	0,72	0,47	0,58
30	0,58	0,68	0,67
31	0,54	0,52	0,55

Teniendo las 30 observaciones

Conocidas

La 31 ava observación perdida es Estimada por la media de las 30 observaciones

[4]

Anexo 8.c

- ✓ p=4, n=30

n	x1	x2	x3	x4
1	0,78	0,66	0,56	0,45
2	0,58	0,35	0,49	0,58
3	0,61	0,52	0,52	0,65
4	0,58	0,47	0,45	0,46
5	0,44	0,65	0,49	0,62
6	0,5	0,43	0,33	0,52
7	0,45	0,56	0,82	0,57
8	0,35	0,56	0,51	0,49
9	0,55	0,4	0,29	0,55
10	0,64	0,38	0,53	0,43
11	0,68	0,55	0,5	0,62
12	0,51	0,64	0,64	0,51
13	0,48	0,57	0,54	0,47
14	0,68	0,45	0,74	0,6
15	0,65	0,58	0,67	0,55
16	0,49	0,46	0,6	0,6
17	0,49	0,57	0,61	0,53
18	0,4	0,54	0,5	0,66
19	0,49	0,57	0,4	0,53
20	0,57	0,33	0,58	0,44
21	0,46	0,57	0,48	0,49
22	0,62	0,52	0,66	0,56
23	0,48	0,72	0,55	0,54
24	0,47	0,42	0,45	0,38
25	0,74	0,34	0,48	0,45
26	0,51	0,59	0,82	0,34
27	0,53	0,58	0,57	0,58
28	0,63	0,43	0,49	0,37
29	0,47	0,49	0,59	0,6
30	0,38	0,58	0,55	0,71

Teniendo las 30 observaciones

Conocidas

[4] Considerar a la 31ava observación como el dato perdido

Anexo 8.d

- ✓ p=4,n=31

n	x1	x2	x3	x4
1	0,78	0,66	0,56	0,45
2	0,58	0,35	0,49	0,58
3	0,61	0,52	0,52	0,65
4	0,58	0,47	0,45	0,46
5	0,44	0,65	0,49	0,62
6	0,5	0,43	0,33	0,52
7	0,45	0,56	0,82	0,57
8	0,35	0,56	0,51	0,49
9	0,55	0,4	0,29	0,55
10	0,64	0,38	0,53	0,43
11	0,68	0,55	0,5	0,62
12	0,51	0,64	0,64	0,51
13	0,48	0,57	0,54	0,47
14	0,68	0,45	0,74	0,6
15	0,65	0,58	0,67	0,55
16	0,49	0,46	0,6	0,6
17	0,49	0,57	0,61	0,53
18	0,4	0,54	0,5	0,66
19	0,49	0,57	0,4	0,53
20	0,57	0,33	0,58	0,44
21	0,46	0,57	0,48	0,49
22	0,62	0,52	0,66	0,56
23	0,48	0,72	0,55	0,54
24	0,47	0,42	0,45	0,38
25	0,74	0,34	0,48	0,45
26	0,51	0,59	0,82	0,34
27	0,53	0,58	0,57	0,58
28	0,63	0,43	0,49	0,37
29	0,47	0,49	0,59	0,6
30	0,38	0,58	0,55	0,71
31	0,54	0,52	0,55	0,53

Teniendo las 30 observaciones

Conocidas

La 31 ava observación perdida es Estimada por la media de las 30 observaciones

Anexo 8.e

✓ p=5,n=30

Teniendo las 30 observaciones

Conocidas

n	x1	x2	x3	x4	x5
1	0,46	0,6	0,31	0,42	0,58
2	0,54	0,63	0,41	0,71	0,61
3	0,76	0,41	0,6	0,53	0,75
4	0,52	0,46	0,74	0,4	0,57
5	0,5	0,61	0,62	0,57	0,36
6	0,35	0,57	0,39	0,45	0,45
7	0,5	0,55	0,71	0,67	0,52
8	0,52	0,6	0,96	0,5	0,41
9	0,4	0,37	0,62	0,44	0,46
10	0,51	0,74	0,44	0,71	0,66
11	0,53	0,51	0,69	0,64	0,36
12	0,48	0,58	0,68	0,37	0,36
13	0,68	0,61	0,19	0,54	0,48
14	0,74	0,56	0,46	0,44	0,35
15	0,44	0,39	0,74	0,56	0,57
16	0,67	0,4	0,5	0,74	0,45
17	0,68	0,68	0,57	0,46	0,51
18	0,37	0,46	0,63	0,61	0,51
19	0,65	0,63	0,57	0,52	0,35
20	0,69	0,47	0,33	0,55	0,53
21	0,71	0,56	0,35	0,7	0,6
22	0,39	0,63	0,64	0,19	0,25
23	0,65	0,6	0,52	0,46	0,66
24	0,55	0,65	0,79	0,53	0,56
25	0,5	0,51	0,54	0,43	0,7
26	0,72	0,73	0,72	0,45	0,48
27	0,53	0,73	0,32	0,6	0,63
28	0,68	0,77	0,49	0,6	0,54
29	0,45	0,61	0,46	0,49	0,67
30	0,42	0,5	0,57	0,43	0,5

Anexo 8.f

- ✓ p=5, n=31

n	x1	x2	x3	x4	x5
1	0.46	0.6	0.31	0.42	0.58
2	0.54	0.63	0.41	0.71	0.61
3	0.76	0.41	0.6	0.53	0.75
4	0.52	0.46	0.74	0.4	0.57
5	0.5	0.61	0.62	0.57	0.36
6	0.35	0.57	0.39	0.45	0.45
7	0.5	0.55	0.71	0.67	0.52
8	0.52	0.6	0.96	0.5	0.41
9	0.4	0.37	0.62	0.44	0.46
10	0.51	0.74	0.44	0.71	0.66
11	0.53	0.51	0.69	0.64	0.36
12	0.48	0.58	0.68	0.37	0.36
13	0.68	0.61	0.19	0.54	0.48
14	0.74	0.56	0.46	0.44	0.35
15	0.44	0.39	0.74	0.56	0.57
16	0.67	0.4	0.5	0.74	0.45
17	0.68	0.68	0.57	0.46	0.51
18	0.37	0.46	0.63	0.61	0.51
19	0.65	0.63	0.57	0.52	0.35
20	0.69	0.47	0.33	0.55	0.53
21	0.71	0.56	0.35	0.7	0.6
22	0.39	0.63	0.64	0.19	0.25
23	0.65	0.6	0.52	0.46	0.66
24	0.55	0.65	0.79	0.53	0.56
25	0.5	0.51	0.54	0.43	0.7
26	0.72	0.73	0.72	0.45	0.48
27	0.53	0.73	0.32	0.6	0.63
28	0,68	0,77	0,49	0,6	0,54
29	0,45	0,61	0,46	0,49	0,67
30	0,42	0,5	0,57	0,43	0,5
31	**0,553**	**0,570**	**0,552**	**0,523**	**0,514**

Teniendo las 30 observaciones Conocidas

La 31 ava observación perdida es Estimada por la media de las 30 observaciones

Anexo 8.g

✓ p=6, n=30

✓

Teniendo las 30 observaciones Conocidas

n	x1	x2	x3	x4	x5	x6
1	0,4886	0,5338	0,3542	0,6104	0,3504	0,4340
2	0,7736	0,4866	0,7296	0,5337	0,4848	0,5453
3	0,6305	0,5555	0,4466	0,7822	0,5143	0,4740
4	0,3880	0,7192	0,5650	0,3389	0,4643	0,4768
5	0,4548	0,6276	0,5167	0,3954	0,4485	0,6559
6	0,6175	0,4676	0,5659	0,4472	0,6675	0,4541
7	0,4035	0,7038	0,5153	0,5199	0,3313	0,6229
8	0,3766	0,3759	0,5752	0,5270	0,5876	0,7114
9	0,5234	0,5498	0,6700	0,4565	0,5732	0,4042
10	0,6168	0,4657	0,5819	0,5585	0,6209	0,5563
11	0,6539	0,6668	0,4365	0,5693	0,4539	0,3990
12	0,6114	0,5206	0,4326	0,6742	0,7028	0,6612
13	0,5228	0,6871	0,6522	0,4473	0,7156	0,5941
14	0,5786	0,4338	0,5232	0,3983	0,4173	0,5146
15	0,4107	0,6972	0,5111	0,3877	0,4676	0,1703
16	0,4497	0,6211	0,5710	0,6299	0,4993	0,3872
17	0,6244	0,5709	0,5658	0,7034	0,7184	0,2967
18	0,4127	0,5044	0,4050	0,3197	0,6007	0,4653
19	0,7591	0,6041	0,4968	0,5711	0,6550	0,4718
20	0,4172	0,5079	0,5060	0,7828	0,5468	0,1585
21	0,5854	0,4738	0,4274	0,5875	0,5082	0,6876
22	0,4750	0,6350	0,3664	0,6180	0,5896	0,6750
23	0,7734	0,5383	0,6167	0,5097	0,4785	0,5032
24	0,2672	0,6619	0,5871	0,6519	0,3806	0,3655
25	0,8193	0,6774	0,3956	0,4371	0,4508	0,5361
26	0,7483	0,4070	0,5544	0,4610	0,7548	0,7190
27	0,3616	0,4347	0,5248	0,8160	0,4729	0,4238
28	0,5450	0,7474	0,3798	0,6593	0,3168	0,3623
29	0,5235	0,4897	0,4435	0,4594	0,4320	0,4559
30	0,5927	0,7212	0,5527	0,5716	0,6285	0,7371

Anexo 8.h

- ✓ p=6, n=31

n	x1	x2	x3	x4	x5	x6
1	0,4886142	0,5337506	0,3541915	0,6103654	0,3504111	0,4340046
2	0,7736144	0,4865765	0,729646	0,5337332	0,4848171	0,5452937
3	0,6304504	0,5554979	0,4466452	0,7822206	0,5142625	0,4740386
4	0,3879934	0,7191606	0,5649692	0,3389283	0,4643456	0,4768407
5	0,4547716	0,6276085	0,5166777	0,3953504	0,448541	0,6559155
6	0,6175251	0,4675745	0,5658839	0,4471977	0,6674924	0,4541213
7	0,4035043	0,7037604	0,5153088	0,5199313	0,331254	0,6229039
8	0,3765919	0,3759354	0,5751985	0,5269691	0,5875503	0,7113699
9	0,5234431	0,5497723	0,6699981	0,4565141	0,5732004	0,4041814
10	0,6168435	0,4657285	0,5819458	0,5585427	0,6208897	0,5562941
11	0,653892	0,6668173	0,4365289	0,569329	0,4539067	0,3990335
12	0,6114023	0,5205916	0,4325764	0,6742272	0,702803	0,6612412
13	0,5227691	0,6870757	0,6522288	0,447341	0,7156444	0,5941192
14	0,5786245	0,4338196	0,5232336	0,3982709	0,4172737	0,5145694
15	0,4106974	0,6971727	0,5110837	0,3877498	0,4675794	0,1703174
16	0,4497023	0,6210827	0,5709885	0,6299067	0,4992877	0,3872496
17	0,6244204	0,5709344	0,5657509	0,703394	0,7184167	0,2966985
18	0,4127261	0,5043992	0,4049599	0,3197316	0,6007323	0,4652602
19	0,7591304	0,6041022	0,4967868	0,5711474	0,6549699	0,4717833
20	0,4171899	0,5079495	0,5059841	0,7828039	0,5468209	0,1585055
21	0,5853802	0,4738182	0,4273606	0,5874847	0,5081952	0,687579
22	0,474962	0,6349868	0,3663833	0,617963	0,5896143	0,6749979
23	0,7733513	0,5383364	0,6166941	0,5097495	0,4784828	0,5032286
24	0,2671683	0,6619272	0,5871258	0,6519409	0,380647	0,3654507
25	0,8193332	0,6773594	0,3956426	0,4371367	0,450835	0,5361017
26	0,7483113	0,4070279	0,5543666	0,4609636	0,7548065	0,7190297
27	0,3615501	0,4346896	0,5248207	0,8159604	0,4728657	0,4238363
28	0,5449807	0,747407	0,3798434	0,6593065	0,3168134	0,3623163
29	0,5234648	0,4897144	0,4435146	0,4593885	0,4320106	0,4558665
30	0,5926733	0,7211687	0,5526938	0,5715636	0,6284772	0,7371137
31	0,54683605	0,56952486	0,51563439	0,54750372	0,52776488	0,49730873

Teniendo las 30 observaciones Conocidas

La 31 ava observación perdida es Estimada por la media de las 30 observaciones

ANEXO B :SIMULACIONES EN EL R CON 3 VARIABLES Y ACP

```
#Definimos el vector de medias "mu" y la matriz de covarianza "SIGMA"

mu=c(0.5387, 8.3466, 1.1857) #vector de medias

SIGMA=matrix(c(0.0146,   0.1113,   0.0258,

                0.1113,   5.6895,   1.0838,

                0.0258,   1.0838,   1.0060), nr=3, nc=3, byrow=T)#matriz de covarianza

#Proceso de instalación de la librería que se va utilizar para la generación de las variables normales multivariables

    install.packages("mnormt")  #La librería "mnormt" nos será de ayuda para poder generar vectores aleatorios de una distribución normal multivariada

    library(mnormt)                  #A continuación "cargaremos" la librería "mnormt"


#Proceso de generación, visulización y resumen descriptivo de las variables normales multivariables

    DATA=rmnorm(5000, mu, SIGMA) #Generaremos 5 mil observaciones con un vector de  medias "mu" y una matriz de covarianza "SIGMA"

DATA=cbind(sample(DATA[,1],size=30,replace=TRUE),sample(DATA[,2],size=30,replace=TRUE),sample(DATA[,3],size=30,replace=TRUE)) #Se extrae 30 observaciones al azar con reemplazo

    DATA=matrix(DATA,nr=30,nc=3,dimnames=list(c(),c("x1","x2","x3")))#Con  la  función "matrix" podremos convertir a la


 #variable DATA en una clase "matrix" con 30 fila y 3 columnas adicionalmente asignaremos nombre a las columnas x1,x2 y x3

DATA                             #Visualizaremos los elementos de la variable DATA

summary(DATA)                    #Con  la  función  "summary"  obtendremos  algunos estadísticos descriptivos de la matriz DATA, tales como el promedio, varianza, etc.

#Obtención de los promedios de cada variable y obtención de la matriz con 31 observaciones así como el resumen descriptivo de la misma


    mu1=mean(DATA[1:30,1])  #Obtendremos el promedio aritmético de la variable x1 (primera columna) obteniendo la variable "mu1"
```

mu2=mean(DATA[1:30,2]) #Obtendremos el promedio aritmético de la variable x2 (segunda columna) obteniendo la variable "mu2"

mu3=mean(DATA[1:30,3]) #Obtendremos el promedio aritmético de la variable x3 (tercera columna) obteniendo la variable "mu3"

mu=c(mu1,mu2,mu3) #Agruparemos los vectores de medias en una sola matriz fila, que llamaremos "mu" y esto es posible por la función "c"

Y=rbind(DATA,mu) #Agruparemos la matriz de medias "mu" con los elementos de la variable DATA y de esa forma obtenemos la matriz Y con 31 observaciones

Y #Visualizaremos los elementos de la variable Y

summary(Y) #Con la función "summary" obtendremos algunos estadísticos descriptivos de la matriz Y

#Obtención de la matriz de correlaciones y autovalores haciendo uso de la matriz de correlación

SIGMA2=cor(DATA) #Determinamos la matriz de correlaciones de las observaciones simuladas y guardadas en la variable DATA

SIGMA2 #Visualizaremos la matriz de correlaciones

eigen(SIGMA2) #Con la función "eigen" obtenemos los autovalores de la matriz de correlaciones SIGMA2

#Determinando las componentes principales con los 30 datos conocidos usando la matriz de correlaciones

summary(pc.cr <- princomp(DATA, cor = TRUE)) #Determinamos un resumen descriptivo de las componentes principales de las 30 observaciones conocidas

loadings(pc.cr)

#Determinando las componentes principales con los 31 datos usando la matriz de correlaciones

mu1=mean(DATA[1:30,1])#Obtendremos el promedio aritmético de la variable x1 (primera columna) obteniendo la variable "mu1"

mu2=mean(DATA[1:30,2])#Obtendremos el promedio aritmético de la variable x2 (segunda columna) obteniendo la variable "mu2"

mu3=mean(DATA[1:30,3])#Obtendremos el promedio aritmético de la variable x3 (tercera columna) obteniendo la variable "mu3"

mu=c(mu1,mu2,mu3) #Agruparemos los vectores de medias en una sola matriz fila, que llamaremos "mu" y esto es posible por la función "c"

```
Y=rbind(DATA,mu)        #Agruparemos la matriz de medias "mu" con los elementos de la variable DATA y de esa forma obtenemos la matriz Y con 31 observaciones

summary(Y)      #Con la función "summary" obtendremos algunos estadísticos descriptivos de la matriz Y

SIGMA3=cor(Y)           #Determinamos la matriz de correlaciones con la variable Y (tiene 31 observaciones)

SIGMA3                  #Visualizaremos la matriz de correlaciones

eigen(SIGMA3)           #Con la función "eigen" obtenemos los autovalores de la matriz de correlaciones SIGMA3

summary(pc.cr <- princomp(Y, cor = TRUE)) #Determinamos un resumen descriptivo de las componentes principales de las 31 observaciones conocidas

loadings(pc.cr)


#Hallando los autovalores y autovectores con la matriz de covarianzas insesgada

SIGMA4=cov(DATA)        #La función "cov" calculará matriz de covarianzas de las 30 observaciones conocidas

SIGMA4                  #Visualizaremos la matriz de covarianza

eigen(SIGMA4)           #Los autovalores de la matriz de covarianzas de las 30 observaciones conocidas


#Determinando las componentes principales con los 30 datos conocidos usando la matriz de covarianzas insesgada

summary(pc.cr <- princomp(DATA, cor = FALSE) )#Determinamos un resumen descriptivo de las componentes principales de las 30 observaciones conocidas

loadings(pc.cr)


#Determinando las componentes principales con los 31 datos usando la matriz de covarianzas insesgada

mu1=mean(DATA[1:30,1]) #Obtendremos el promedio aritmético de la variable x1 (primera columna) obteniendo la variable "mu1"

mu2=mean(DATA[1:30,2]) #Obtendremos el promedio aritmético de la variable x2 (segunda columna) obteniendo la variable "mu2"
```

mu3=mean(DATA[1:30,3]) #Obtendremos el promedio aritmético de la variable x3 (tercera columna) obteniendo la variable "mu3"

mu=c(mu1,mu2,mu3) #Agruparemos los vectores de medias en una sola matriz fila, que llamaremos "mu" y esto es posible por la función "c"

Y=rbind(DATA,mu) #Agruparemos la matriz de medias "mu" con los elementos de la variable DATA y de esa forma obtenemos la matriz Y con 31 observaciones

Y #Visualizaremos los elementos de la variable Y

summary(Y) #Con la función "summary" obtendremos algunos estadísticos descriptivos de la matriz Y

SIGMA5=cov(Y) #La función "cov" calculará matriz de covarianzas de las 31 observaciones conocidas

SIGMA5 #Visualizaremos la matriz de covarianza

eigen(SIGMA5) #Los autovalores de la matriz de covarianzas de las 31 observaciones conocidas

summary(pc.cr <- princomp(Y, cor = FALSE)) #Determinamos un resumen descriptivo de las componentes principales de las 31 observaciones conocidas

loadings(pc.cr)

#Hallando los autovalores y autovectores con la matriz de covarianzas sesgada

SIGMA6=(29/30)*cov(DATA) #La variable SIGMA6 representa la matriz de covarianzas sesgada de las 30 observaciones conocidas

SIGMA6 #Visualizaremos la matriz de covarianza

eigen(SIGMA6) #Los autovalores de la matriz de covarianzas de las 30 observaciones conocidas

#Hallando los autovalores y autovectores con la matriz de covarianzas sesgada con los 31 datos

SIGMA7=(29/30)*cov(Y) #La variable SIGMA7 representa la matriz de covarianzas sesgada de las 31 observaciones conocidas

SIGMA7 #Visualizaremos la matriz de covarianza

eigen(SIGMA7) #Los autovalores de la matriz de covarianzas sesgada de las 31 observaciones

ANEXO C :SIMULACIONES EN EL R CON 4 VARIABLES Y ACP

```
#Definimos el vector de medias "mu" y la matriz de covarianza "SIGMA"

    mu=c(0.5387, 8.3466, 1.1857 ,0.9611)

    SIGMA=matrix(c(0.0146,  0.1113,  0.0258,  0.0083,

        0.1113,  5.6895,  1.0838,  0.1143,

        0.0258,  1.0838,  1.0060,  0.1091,

        0.0083,  0.1143,  0.1091,  0.1276), nr=4, nc=4, byrow=T) #matriz de
covarianza

#Proceso de instalación de la librería que se va utilizar para la generación de las variables
normales multivariables

    install.packages("mnormt") #La librería "mnormt" nos será de ayuda para poder
generar vectores aleatorios de una distribución normal multivariada

    library(mnormt)          #A continuación "cargaremos" la librería "mnormt"


#Proceso de generación, visulización y resumen descriptivo de las variables normales
multivariables

    DATA=rmnorm(5000, mu, SIGMA) ) #Generaremos 5 mil observaciones con un
vector de  medias "mu" y una matriz de covarianza "SIGMA"

    DATA=cbind(sample(DATA[,1],size=30,replace=TRUE),sample(DATA[,2],size=30,rep
lace=TRUE),sample(DATA[,3],size=30,replace=TRUE),sample(DATA[,4],size=30,replace=T
RUE)) #Se extrae 30 observaciones al azar con reemplazo

    DATA=matrix(DATA,nr=30,nc=4,dimnames=list(c(),c("x1","x2","x3","x4"))) #Se
convierte la clase "numerico" a "matriz" con la función matrix

DATA                         #Visualizaremos los elementos de la variable DATA

    summary(DATA)            #Con la función "summary" obtendremos algunos
estadísticos descriptivos de la matriz DATA


#Obtención de los promedios de cada variable y obtención de la matriz con 31
observaciones así como el resumen descriptivo de la misma

    mu1=mean(DATA[1:30,1]) #Obtendremos el promedio aritmético de la variable x1
(primera columna) obteniendo la variable "mu1"

    mu2=mean(DATA[1:30,2]) #Obtendremos el promedio aritmético de la variable x2
(segunda columna) obteniendo la variable "mu2"
```

mu3=mean(DATA[1:30,3]) #Obtendremos el promedio aritmético de la variable x3 (tercera columna) obteniendo la variable "mu3"

mu4=mean(DATA[1:30,4]) #Obtendremos el promedio aritmético de la variable x4 (cuarta columna) obteniendo la variable "mu4"

mu=c(mu1,mu2,mu3,mu4) #Agruparemos los vectores de medias en una sola matriz fila, que llamaremos "mu" y esto es posible por la función "c"

Y=rbind(DATA,mu) #Agruparemos la matriz de medias "mu" con los elementos de la variable DATA y de esa forma obtenemos la matriz Y con 31 observaciones

Y #Visualizaremos los elementos de la variable Y

summary(Y) #Con la función "summary" obtendremos algunos estadísticos descriptivos de la matriz Y

#Obtención de la matriz de correlaciones y autovalores haciendo uso de la matriz de correlación

SIGMA2=cor(DATA) #Determinamos la matriz de correlaciones de las observaciones simuladas y guardadas en la variable DATA

SIGMA2 #Visualizaremos la matriz de correlaciones

eigen(SIGMA2) #Con la función "eigen" obtenemos los autovalores de la matriz de correlaciones SIGMA2

#Determinando las componentes principales con los 30 datos conocidos usando la matriz de correlaciones

summary(pc.cr <- princomp(DATA, cor = TRUE)) #Determinamos un resumen descriptivo de las componentes principales de las 30 observaciones conocidas

loadings(pc.cr)

#Determinando las componentes principales con los 31 datos usando la matriz de correlaciones

mu1=mean(DATA[1:30,1]) #Obtendremos el promedio aritmético de la variable x1 (primera columna) obteniendo la variable "mu1"

mu2=mean(DATA[1:30,2]) #Obtendremos el promedio aritmético de la variable x2 (segunda columna) obteniendo la variable "mu2"

mu3=mean(DATA[1:30,3]) #Obtendremos el promedio aritmético de la variable x3 (tercera columna) obteniendo la variable "mu3"

mu4=mean(DATA[1:30,4]) #Obtendremos el promedio aritmético de la variable x4 (cuarta columna) obteniendo la variable "mu4"

mu=c(mu1,mu2,mu3,mu4) #Agruparemos los vectores de medias en una sola matriz fila, que llamaremos "mu" y esto es posible por la función "c"

Y=rbind(DATA,mu) #Agruparemos la matriz de medias "mu" con los elementos de la variable DATA y de esa forma obtenemos la matriz Y con 31 observaciones

summary(Y) #Con la función "summary" obtendremos algunos estadísticos descriptivos de la matriz Y

SIGMA3=cor(Y) #Determinamos la matriz de correlaciones con la variable Y (tiene 31 observaciones)

SIGMA3 #Visualizaremos la matriz de correlaciones

eigen(SIGMA3) #Con la función "eigen" obtenemos los autovalores de la matriz de correlaciones SIGMA3

summary(pc.cr <- princomp(Y, cor = TRUE)) #Determinamos un resumen descriptivo de las componentes principales de las 31 observaciones conocidas

loadings(pc.cr)

#Hallando los autovalores y autovectores con la matriz de covarianzas insesgada

SIGMA4=cov(DATA #La función "cov" calculará matriz de covarianzas de las 30 observaciones conocidas

SIGMA4 #Visualizaremos la matriz de covarianza

eigen(SIGMA4) #Los autovalores de la matriz de covarianzas de las 30 observaciones conocidas

#Determinando las componentes principales con los 30 datos conocidos usando la matriz de covarianzas insesgada

summary(pc.cr <- princomp(DATA, cor = FALSE)) #Determinamos un resumen descriptivo de las componentes principales de las 30 observaciones conocidas

loadings(pc.cr)

#Determinando las componentes principales con los 31 datos usando la matriz de covarianzas insesgada

mu1=mean(DATA[1:30,1]) #Obtendremos el promedio aritmético de la variable x1 (primera columna) obteniendo la variable "mu1"

mu2=mean(DATA[1:30,2]) #Obtendremos el promedio aritmético de la variable x2 (segunda columna) obteniendo la variable "mu2"

mu3=mean(DATA[1:30,3]) #Obtendremos el promedio aritmético de la variable x3 (tercera columna) obteniendo la variable "mu3"

mu4=mean(DATA[1:30,4]) #Obtendremos el promedio aritmético de la variable x4 (cuarta columna) obteniendo la variable "mu4"

mu=c(mu1,mu2,mu3,mu4) #Agruparemos los vectores de medias en una sola matriz fila, que llamaremos "mu" y esto es posible por la función "c"

Y=rbind(DATA,mu) #Agruparemos la matriz de medias "mu" con los elementos de la variable DATA y de esa forma obtenemos la matriz Y con 31 observaciones

Y #Visualizaremos los elementos de la variable Y

summary(Y) #Con la función "summary" obtendremos algunos estadísticos descriptivos de la matriz Y

SIGMA5=cov(Y) #La función "cov" calculará matriz de covarianzas de las 31 observaciones conocidas

SIGMA5 #Visualizaremos la matriz de covarianza

eigen(SIGMA5) #Los autovalores de la matriz de covarianzas de las 31 observaciones conocidas

summary(pc.cr <- princomp(Y, cor = FALSE)) #Determinamos un resumen descriptivo de las componentes principales de las 31 observaciones conocidas

loadings(pc.cr)

#Hallando los autovalores y autovectores con la matriz de covarianzas sesgada

SIGMA6=(29/30)*cov(DATA)#La variable SIGMA6 representa la matriz de covarianzas sesgada de las 30 observaciones conocidas

SIGMA6 #Visualizaremos la matriz de covarianza

eigen(SIGMA6) #Los autovalores de la matriz de covarianzas de las 30 observaciones conocidas

#Hallando los autovalores y autovectores con la matriz de covarianzas sesgada con los 31 datos

SIGMA7=(29/30)*cov(Y) #La variable SIGMA7 representa la matriz de covarianzas sesgada de las 31 observaciones conocidas

SIGMA7 #Visualizaremos la matriz de covarianza

```
eigen(SIGMA7)                #Los autovalores de la matriz de covarianzas sesgada de las 31 observaciones
```

ANEXO C :SIMULACIONES EN EL R CON 5 VARIABLES Y ACP

```
#Definimos el vector de medias "mu" y la matriz de covarianza "SIGMA"

mu=c(0.5387, 8.3466, 1.1857 ,0.9611 ,4.2101) #vector de medias

SIGMA=matrix(c(0.0146,  0.1113,  0.0258,  0.0083,  0.0139,
       0.1113,  5.6895,  1.0838,  0.1143,  0.9412,
       0.0258,  1.0838,  1.0060,  0.1091,  0.6846,
       0.0083,  0.1143,  0.1091,  0.1276,  0.1485,
       0.0139,  0.9412,  0.6846,  0.1485,  2.1059), nr=5, nc=5, byrow=T) #matriz de covarianza


#Proceso de instalación de la librería que se va utilizar para la generación de las variables normales multivariables

install.packages("mnormt")  #La librería "mnormt" nos será de ayuda para poder generar vectores aleatorios de una distribución normal multivariada

library(mnormt)             #A continuación "cargaremos" la librería "mnormt"


#Proceso de generación, visulización y resumen descriptivo de las variables normales multivariables

DATA=rmnorm(5000, mu, SIGMA) #Generaremos 5 mil observaciones con un vector de  medias "mu" y una matriz de covarianza "SIGMA"

DATA=cbind(sample(DATA[,1],size=30,replace=TRUE),sample(DATA[,2],size=30,replace=TRUE),sample(DATA[,3],size=30,replace=TRUE),sample(DATA[,4],size=30,replace=TRUE),sample(DATA[,5],size=30,replace=TRUE)) #Se extrae 30 observaciones al azar con reemplazo

DATA=matrix(DATA,nr=30,nc=5,dimnames=list(c(),c("x1","x2","x3","x4","x5","x6"))) #Se convierte la clase "numerico" a "matriz" con la función matrix

DATA                        #Visualizaremos los elementos de la variable DATA

summary(DATA)  #Con la función "summary" obtendremos algunos estadísticos descriptivos de la matriz DATA
```

#Obtención de los promedios de cada variable y obtención de la matriz con 31 observaciones así como el resumen descriptivo de la misma

mu1=mean(DATA[1:30,1]) #Obtendremos el promedio aritmético de la variable x1 (primera columna) obteniendo la variable "mu1"

mu2=mean(DATA[1:30,2]) #Obtendremos el promedio aritmético de la variable x2 (segunda columna) obteniendo la variable "mu2"

mu3=mean(DATA[1:30,3]) #Obtendremos el promedio aritmético de la variable x3 (tercera columna) obteniendo la variable "mu3"

mu4=mean(DATA[1:30,4]) #Obtendremos el promedio aritmético de la variable x4 (cuarta columna) obteniendo la variable "mu4"

mu5=mean(DATA[1:30,5]) #Obtendremos el promedio aritmético de la variable x5 (quinta columna) obteniendo la variable "mu5"

mu=c(mu1,mu2,mu3,mu4,mu5) #Agruparemos los vectores de medias en una sola matriz fila, que llamaremos "mu" y esto es posible por la función "c"

Y=rbind(DATA,mu) #Agruparemos la matriz de medias "mu" con los elementos de la variable DATA y de esa forma obtenemos la matriz Y con 31 observaciones

Y #Visualizaremos los elementos de la variable y

summary(Y) #Con la función "summary" obtendremos algunos estadísticos descriptivos de la matriz Y

#Obtención de la matriz de correlaciones y autovalores haciendo uso de la matriz de correlación

SIGMA2=cor(DATA) #Determinamos la matriz de correlaciones de las observaciones simuladas y guardadas en la variable DATA

SIGMA2 #Visualizaremos la matriz de correlaciones

eigen(SIGMA2) #Con la función "eigen" obtenemos los autovalores de la matriz de correlaciones SIGMA2

#Determinando las componentes principales con los 30 datos conocidos usando la matriz de correlaciones

summary(pc.cr <- princomp(DATA, cor = TRUE)) #Determinamos un resumen descriptivo de las componentes principales de las 30 observaciones conocidas

loadings(pc.cr)

#Determinando las componentes principales con los 31 datos usando la matriz de correlaciones

mu1=mean(DATA[1:30,1]) #Obtendremos el promedio aritmético de la variable x1 (primera columna) obteniendo la variable "mu1"

mu2=mean(DATA[1:30,2]) #Obtendremos el promedio aritmético de la variable x2 (segunda columna) obteniendo la variable "mu2"

mu3=mean(DATA[1:30,3]) #Obtendremos el promedio aritmético de la variable x3 (tercera columna) obteniendo la variable "mu3"

mu4=mean(DATA[1:30,4]) #Obtendremos el promedio aritmético de la variable x4 (cuarta columna) obteniendo la variable "mu4"

mu5=mean(DATA[1:30,5]) #Obtendremos el promedio aritmético de la variable x5 (quinta columna) obteniendo la variable "mu5"

mu=c(mu1,mu2,mu3,mu4,mu5) #Agruparemos los vectores de medias en una sola matriz fila, que llamaremos "mu" y esto es posible por la función "c"

Y=rbind(DATA,mu) #Agruparemos la matriz de medias "mu" con los elementos de la variable DATA y de esa forma obtenemos la matriz Y con 31 observaciones

Y #Visualizaremos los elementos de la variable y

summary(Y) #Con la función "summary" obtendremos algunos estadísticos descriptivos de la matriz Y

SIGMA3=cor(Y) #Determinamos la matriz de correlaciones con la variable Y (tiene 31 observaciones)

SIGMA3 #Visualizaremos la matriz de correlaciones

eigen(SIGMA3) #Con la función "eigen" obtenemos los autovalores de la matriz de correlaciones SIGMA3

summary(pc.cr <- princomp(Y, cor = TRUE)) #Determinamos un resumen descriptivo de las componentes principales de las 31 observaciones conocidas

loadings(pc.cr)

#Hallando los autovalores y autovectores con la matriz de covarianzas insesgada

SIGMA4=cov(DATA #La función "cov" calculará matriz de covarianzas de las 30 observaciones conocidas

SIGMA4 #Visualizaremos la matriz de covarianza

```
    eigen(SIGMA4)          #Los autovalores de la matriz de covarianzas de las 30 observaciones conocidas


#Determinando las componentes principales con los 30 datos conocidos usando la matriz de covarianzas insesgada

    summary(pc.cr <- princomp(DATA, cor = FALSE)) #Determinamos un resumen descriptivo de las componentes principales de las 30 observaciones conocidas

    loadings(pc.cr)

#Determinando las componentes principales con los 31 datos usando la matriz de covarianzas insesgada

    mu1=mean(DATA[1:30,1])          #Obtendremos el promedio aritmético de la variable x1 (primera columna) obteniendo la variable "mu1"

    mu2=mean(DATA[1:30,2])          #Obtendremos el promedio aritmético de la variable x2 (segunda columna) obteniendo la variable "mu2"

    mu3=mean(DATA[1:30,3])          #Obtendremos el promedio aritmético de la variable x3 (tercera columna) obteniendo la variable "mu3"

    mu4=mean(DATA[1:30,4])          #Obtendremos el promedio aritmético de la variable x4 (cuarta columna) obteniendo la variable "mu4"

    mu5=mean(DATA[1:30,5])  #Obtendremos el promedio aritmético de la variable x5 (quinta columna) obteniendo la variable "mu5"


    mu=c(mu1,mu2,mu3,mu4,mu5) #Agruparemos los vectores de medias en una sola matriz fila, que llamaremos "mu" y esto es posible por la función "c"

    Y=rbind(DATA,mu)                #Agruparemos la matriz de medias "mu" con los elementos de la variable DATA y de esa forma obtenemos la matriz Y con 31 observaciones

    Y                               #Visualizaremos los elementos de la variable y

    summary(Y)                      #Con la función "summary" obtendremos algunos estadísticos descriptivos de la matriz Y

    SIGMA5=cov(Y)          #La función "cov" calculará matriz de covarianzas de las 31 observaciones conocidas

    SIGMA5                 #Visualizaremos la matriz de covarianza

    eigen(SIGMA5)          #Los autovalores de la matriz de covarianzas de las 31 observaciones conocidas

    summary(pc.cr <- princomp(Y, cor = FALSE)) #Determinamos un resumen descriptivo de las componentes principales de las 31 observaciones conocidas
```

```
loadings(pc.cr)


#Hallando los autovalores y autovectores con la matriz de covarianzas sesgada

SIGMA6=(29/30)*cov(DATA)#La variable SIGMA6 representa la matriz de covarianzas sesgada de las 30 observaciones conocidas

SIGMA6                  #Visualizaremos la matriz de covarianza

eigen(SIGMA6)           #Los autovalores de la matriz de covarianzas de las 30 observaciones conocidas

#Hallando los autovalores y autovectores con la matriz de covarianzas sesgada con los 31 datos

SIGMA7=(29/30)*cov(Y)   #La variable SIGMA7 representa la matriz de covarianzas sesgada de las 31 observaciones conocidas

SIGMA7                  #Visualizaremos la matriz de covarianza

eigen(SIGMA7)           #Los autovalores de la matriz de covarianzas sesgada de las 31 observaciones
```

ANEXO D :SIMULACIONES EN EL R CON 6 VARIABLES Y ACP

```
#Definimos el vector de medias "mu" y la matriz de covarianza "SIGMA"

mu=c(0.5387, 8.3466, 1.1857 ,0.9611 ,4.2101 ,2.8149) #vector de medias

SIGMA=matrix(c(0.0146,  0.1113,  0.0258,  0.0083,  0.0139,  0.0025,
       0.1113,  5.6895,  1.0838,  0.1143,  0.9412,  0.1515,
       0.0258,  1.0838,  1.0060,  0.1091,  0.6846,  0.1577,
       0.0083,  0.1143,  0.1091,  0.1276,  0.1485,  0.0085,
       0.0139,  0.9412,  0.6846,  0.1485,  2.1059,  0.1564,
       0.0025,  0.1515,  0.1577,  0.0085,  0.1564,  0.1029), nr=6, nc=6, byrow=T) #matriz de covarianza


#Proceso de instalación de la librería que se va utilizar para la generación de las variables normales multivariables

install.packages("mnormt")  #La librería "mnormt" nos será de ayuda para poder generar vectores aleatorios de una distribución normal multivariada

library(mnormt)             #A continuación "cargaremos" la librería "mnormt"
```

#Proceso de generación, visulización y resumen descriptivo de las variables normales multivariables

DATA=rmnorm(5000, mu, SIGMA) #Generaremos 5 mil observaciones con un vector de medias "mu" y una matriz de covarianza "SIGMA"

DATA=cbind(sample(DATA[,1],size=30,replace=TRUE),sample(DATA[,2],size=30,replace=TRUE),sample(DATA[,3],size=30,replace=TRUE),sample(DATA[,4],size=30,replace=TRUE),sample(DATA[,5],size=30,replace=TRUE),sample(DATA[,6],size=30,replace=TRUE)) #Se extrae 30 observaciones al azar con reemplazo

DATA=matrix(DATA,nr=30,nc=6,dimnames=list(c(),c("x1","x2","x3","x4","x5","x6"))) #Se convierte la clase "numerico" a "matriz" con la función matrix

DATA #Visualizaremos los elementos de la variable DATA

summary(DATA) #Con la función "summary" obtendremos algunos estadísticos descriptivos de la matriz DATA

#Obtención de los promedios de cada variable y obtención de la matriz con 31 observaciones así como el resumen descriptivo de la misma

mu1=mean(DATA[1:30,1]) #Obtendremos el promedio aritmético de la variable x1 (primera columna) obteniendo la variable "mu1"

mu2=mean(DATA[1:30,2]) #Obtendremos el promedio aritmético de la variable x2 (segunda columna) obteniendo la variable "mu2"

mu3=mean(DATA[1:30,3]) #Obtendremos el promedio aritmético de la variable x3 (tercera columna) obteniendo la variable "mu3"

mu4=mean(DATA[1:30,4]) #Obtendremos el promedio aritmético de la variable x4 (cuarta columna) obteniendo la variable "mu4"

mu5=mean(DATA[1:30,5]) #Obtendremos el promedio aritmético de la variable x5 (quinta columna) obteniendo la variable "mu5"

mu6=mean(DATA[1:30,6]) #Obtendremos el promedio aritmético de la variable x6 (sexta columna) obteniendo la variable "mu6"

mu=c(mu1,mu2,mu3,mu4,mu5,mu6) #Agruparemos los vectores de medias en una sola matriz fila, que llamaremos "mu" y esto es posible por la función "c"

Y=rbind(DATA,mu) #Agruparemos la matriz de medias "mu" con los elementos de la variable DATA y de esa forma obtenemos la matriz Y con 31 observaciones

Y #Visualizaremos los elementos de la variable y

summary(Y) #Con la función "summary" obtendremos algunos estadísticos descriptivos de la matriz Y

#Obtención de la matriz de correlaciones y autovalores haciendo uso de la matriz de correlación

SIGMA2=cor(DATA) #Determinamos la matriz de correlaciones de las observaciones simuladas y guardadas en la variable DATA

SIGMA2 #Visualizaremos la matriz de correlaciones

eigen(SIGMA2) #Con la función "eigen" obtenemos los autovalores de la matriz de correlaciones SIGMA2

#Determinando las componentes principales con los 30 datos conocidos usando la matriz de correlaciones

summary(pc.cr <- princomp(DATA, cor = TRUE)) #Determinamos un resumen descriptivo de las componentes principales de las 30 observaciones conocidas

loadings(pc.cr)

#Determinando las componentes principales con los 31 datos usando la matriz de correlaciones

mu1=mean(DATA[1:30,1]) #Obtendremos el promedio aritmético de la variable x1 (primera columna) obteniendo la variable "mu1"

mu2=mean(DATA[1:30,2]) #Obtendremos el promedio aritmético de la variable x2 (segunda columna) obteniendo la variable "mu2"

mu3=mean(DATA[1:30,3]) #Obtendremos el promedio aritmético de la variable x3 (tercera columna) obteniendo la variable "mu3"

mu4=mean(DATA[1:30,4]) #Obtendremos el promedio aritmético de la variable x4 (cuarta columna) obteniendo la variable "mu4"

mu5=mean(DATA[1:30,5]) #Obtendremos el promedio aritmético de la variable x5 (quinta columna) obteniendo la variable "mu5"

mu6=mean(DATA[1:30,6]) #Obtendremos el promedio aritmético de la variable x6 (sexta columna) obteniendo la variable "mu6"

mu=c(mu1,mu2,mu3,mu4,mu5,mu6) #Agruparemos los vectores de medias en una sola matriz fila, que llamaremos "mu" y esto es posible por la función "c"

Y=rbind(DATA,mu) #Agruparemos la matriz de medias "mu" con los elementos de la variable DATA y de esa forma obtenemos la matriz Y con 31 observaciones

Y #Visualizaremos los elementos de la variable Y

summary(Y) #Con la función "summary" obtendremos algunos estadísticos descriptivos de la matriz Y

SIGMA3=cor(Y) #Determinamos la matriz de correlaciones con la variable Y (tiene 31 observaciones)

SIGMA3 #Visualizaremos la matriz de correlaciones

eigen(SIGMA3) #Con la función "eigen" obtenemos los autovalores de la matriz de correlaciones SIGMA3

summary(pc.cr <- princomp(Y, cor = TRUE)) #Determinamos un resumen descriptivo de las componentes principales de las 31 observaciones conocidas

loadings(pc.cr)

#Hallando los autovalores y autovectores con la matriz de covarianzas insesgada

SIGMA4=cov(DATA #La función "cov" calculará matriz de covarianzas de las 30 observaciones conocidas

SIGMA4 #Visualizaremos la matriz de covarianza

eigen(SIGMA4) #Los autovalores de la matriz de covarianzas de las 30 observaciones conocidas

#Determinando las componentes principales con los 30 datos conocidos usando la matriz de covarianzas insesgada

summary(pc.cr <- princomp(DATA, cor = FALSE)) #Determinamos un resumen descriptivo de las componentes principales de las 30 observaciones conocidas

loadings(pc.cr)

#Determinando las componentes principales con los 31 datos usando la matriz de covarianzas insesgada

mu1=mean(DATA[1:30,1]) #Obtendremos el promedio aritmético de la variable x1 (primera columna) obteniendo la variable "mu1"

mu2=mean(DATA[1:30,2]) #Obtendremos el promedio aritmético de la variable x2 (segunda columna) obteniendo la variable "mu2"

mu3=mean(DATA[1:30,3]) #Obtendremos el promedio aritmético de la variable x3 (tercera columna) obteniendo la variable "mu3"

mu4=mean(DATA[1:30,4]) #Obtendremos el promedio aritmético de la variable x4 (cuarta columna) obteniendo la variable "mu4"

mu5=mean(DATA[1:30,5]) #Obtendremos el promedio aritmético de la variable x5 (quinta columna) obteniendo la variable "mu5"

```
mu6=mean(DATA[1:30,6])        #Obtendremos el promedio aritmético de la variable x6 (sexta columna) obteniendo la variable "mu6"

mu=c(mu1,mu2,mu3,mu4,mu5,mu6) #Agruparemos los vectores de medias en una sola matriz fila, que llamaremos "mu" y esto es posible por la función "c"

Y=rbind(DATA,mu)              #Agruparemos la matriz de medias "mu" con los elementos de la variable DATA y de esa forma obtenemos la matriz Y con 31 observaciones

Y                             #Visualizaremos los elementos de la variable y

summary(Y)                    #Con la función "summary" obtendremos algunos estadísticos descriptivos de la matriz Y

SIGMA5=cov(Y)                 #La función "cov" calculará matriz de covarianzas de las 31 observaciones conocidas

SIGMA5                        #Visualizaremos la matriz de covarianza

eigen(SIGMA5)                 #Los autovalores de la matriz de covarianzas de las 31 observaciones conocidas

summary(pc.cr <- princomp(Y, cor = FALSE)) #Determinamos un resumen descriptivo de las componentes principales de las 31 observaciones conocidas

loadings(pc.cr)

#Hallando los autovalores y autovectores con la matriz de covarianzas sesgada

SIGMA6=(29/30)*cov(DATA)#La variable SIGMA6 representa la matriz de covarianzas sesgada de las 30 observaciones conocidas

SIGMA6                        #Visualizaremos la matriz de covarianza

eigen(SIGMA6)                 #Los autovalores de la matriz de covarianzas de las 30 observaciones conocidas

#Hallando los autovalores y autovectores con la matriz de covarianzas sesgada con los 31 datos

SIGMA7=(29/30)*cov(Y)  #La variable SIGMA7 representa la matriz de covarianzas sesgada de las 31 observaciones conocidas

SIGMA7                        #Visualizaremos la matriz de covarianza

eigen(SIGMA7)                 #Los autovalores de la matriz de covarianzas sesgada de las 31 observaciones
```